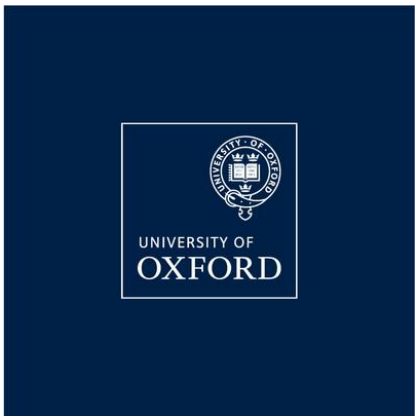
Analytical Pathways for Long-Read Sequencing in Neuropathology



Hannah Brooks

Linacre College

University of Oxford

Supervisors

Dr Olaf Ansorge

Dr Casmir Turnquist

A thesis submitted for the degree of *Doctor of Philosophy – Clinical Neurosciences*

Michaelmas 2025

Abstract

Hannah Brooks, Linacre College

Rare solid neoplasms of the craniospinal axis are difficult to diagnose due to tumour heterogeneity, and limited understanding of their molecular and genetic drivers has restricted the development of targeted therapies, resulting in reliance on surgery and radiotherapy. Consequently, the prospect of 'precision oncology' – defined as the right treatment for the right tumour in the right person at the right time – remains unfulfilled. This thesis explores how advances in long-read Oxford Nanopore Technologies (ONT) sequencing could achieve rapid histomolecular 'precision diagnostics' and help dissect the biology of a rare tumour called mesenchymal chondrosarcoma (MCS). The premise of the work is that tissue diagnostics and research in rare neoplasms must be integrated and barriers to access to cutting edge technology removed. This means that novel tissue pathways must be developed and validated in the research setting before they can be implemented as standard of care. Therefore, I set out to examine three related questions:

1. How can we optimally preserve tissue morphology and molecular integrity for rapid ONT diagnostics and research?

2. Can ONT sequencing provide molecular data that allows us to make an integrated histomolecular diagnosis in hours rather than weeks?

3. Can ONT resolve the fusion-gene architecture of MCS at bulk and single cell level to illuminate its biology?

I approached these questions with a combination of histological and emerging ONT wet-lab and bioinformatics technologies.

1. I found that a novel approach to rapid tissue freezing without embedding medium or other toxic chemicals is superior to any other tissue preservation technique, as it does not induce any detrimental modifications and still preserves cytoarchitecture.

2. I determined that ONT sequencing of rare neoplasms from both biopsy and postmortem tissues could provide equivalent and, in some cases, superior molecular diagnostic data in a fraction of the time compared to standard-of-care testing.

3. I established pipelines for comparative DNA and RNA ONT long-read sequencing of MCS at bulk and single cell level which allowed me to identify and characterise the pathognomonic *HEY1::NCOA2* fusion in its genomic and cellular context.

In conclusion, my thesis establishes the foundations for a novel, rapid, non-toxic and cost-effective tissue pathway for near-patient precision oncology for people with rare neoplasms which is suitable for future implementation in the NHS. Further, my work contributes the first bulk and single cell ONT long-read dataset of MCS, an ultra-rare tumour whose biology remains poorly understood.

Acknowledgements

First and foremost, I would like to thank my supervisors, Professor Olaf Ansorge and Dr Casmir Turnquist for giving me the opportunity to complete this project. Your guidance and expertise were instrumental in the completion of this thesis.

I would like to extend a special thanks to Ebony Cave, whose bioinformatics expertise was vital to the success of this project. The bioinformatics work was a crucial component of this thesis, and I truly could not have completed it without your invaluable contributions. I would also like to thank Andrew Lee for your technical expertise and performing the GEM generation and barcoding steps of the single-nuclei sequencing protocol.

I would also like to thank my colleague, Carolyn Sloan, for your constant support and always being there when I needed you. I have loved working with you and truly appreciate all your help.

Thank you to my parents, Glyn and Sue Brooks and my sister, Emily. From the very beginning you have supported me in every way possible. Your encouragement and belief in me have been the driving force of everything I have accomplished, and I will be forever grateful. I am incredibly fortunate to have you as my family.

To my husband, Jono. Words cannot express how thankful I am for your endless love, patience and support throughout this journey. You have helped and encouraged me, even during the most difficult times. Thank you for always being there and for everything you have done to make this possible. I couldn't have completed this without you.

Lastly, to my daughter, Poppy. Your laughter and joy have been a constant source of motivation throughout this challenge. Even though you are too little to read this now, I hope that one day you understand how much you helped me. I dedicate this achievement to you.

Declaration

I, Hannah Brooks, hereby declare that the work on which this thesis is based was performed by myself, unless otherwise listed in the text as follows: long-read and short-read whole genome sequencing bioinformatic pipeline and single-nuclei whole genome sequencing bioinformatics pipeline was performed by Dr Ebony cave; GEM generation and barcoding was performed by Andrew Lee.

Contents

List of Figures

List of Tables

Abbreviations

5mC – 5-methylcytosine

5hmC – 5-hydroxymethylcytosine

ATRT – Atypical Teratoid Rhabdoid Tumour

CNS – Central Nervous System

CNV – Copy Number Variant

CpG - Cytosine-Guanine dinucleotides

DIPG – Diffuse Intrinsic Pontine Glioma

EM – Electron Microscopy

eRMS – Embryonal Rhabdomyosarcoma

FFPE – Formalin Fixed Paraffin Embedded

FISH – Fluorescence In Situ Hybridization

GBM – Glioblastoma

GEMs – Gel Beads-in-Emulsion

H&E – Haematoxylin and Eosin

HAC – High-Accuracy Model

HMW – High Molecular Weight

IHC – Immunohistochemistry

LCH – Langerhans Cell Histiocytosis

LGG – Low Grade Glioma

LR – Long-Read Sequencing

MCS – Mesenchymal Chondrosarcoma

MCC – Mileston Cryoembedding Compound

MNG – Meningioma

MNV – Multi-Nucleotide Variant

nCount – Median UMI's per Cell

nFeature – Median Number of Genes per Cell

NGS- Next Generation Sequencing

ONT – Oxford Nanopore Technologies

PCV – Procarbazine, Lomustine and Vincristine Chemotherapy

PCR – Polymerase Chain Reaction

PM – Post-Mortem

RF – Random Forest

ROBIN – Rapid nanOpore Brain Intraoperative classification

RT – Room Temperature

RT-PCR – Reverse Transcriptase-Polymerase Chain Reaction

SARC – RMS – Sarcoma – Rhabdomyosarcoma

SNV – Single-Nucleotide Variant

SoC – Standard of Care

SR – Short-Read Sequencing

SV – Structural Variants

TJ-BM – Tess Jowell BRAIN MATRIX Study

UMI – Unique Molecular Identifiers

VAF – Variant Allele Frequency

WES – Whole Exome Sequencing

WGS – Whole Genome Sequencing

WHO – World Health Organization

Chapter 1 Introduction

1.1 Scope of DPhil project

Over 12,000 people are diagnosed with a tumour of the central nervous system (CNS) each year in the UK (CRUK, 2025). The incidence rate has been predicted to increase 6% over 20 years due to an aging population and advances in diagnostic techniques such as high-resolution neuroimaging (Zhou *et al*, 2025). The most commonly occurring and most malignant primary tumours of the central nervous system are gliomas. Gliomas are classified depending on the presumed cell of origin. Glioblastoma is the most malignant type of glioma with the median survival being only 15 months (Ohgaki and Kleihues, 2005; Thakkar *et al*, 2014; Hanif *et al,* 2017*)*. These tumours are extremely difficult to treat as the location and diffuse growth pattern of the tumours within the brain makes gross total resection impossible. Further, the presence of the blood brain barrier (although compromised in the most malignant gliomas), reduce the efficacy of chemotherapy drugs that successfully work in treating cancers of other organs (Bender, 2018; Han *et al,* 2020). It is for these reasons that brain tumours have been identified as a priority research area by CRUK and the UK government, resulting in the establishment of the Tessa Jowell BRAIN MATRIX study (TJ-BM). This study includes ten centres around the UK and has received funding from The Brain Tumour Charity and Industrial Strategy Challenge Fund for five years. The main aim of this project is to develop infrastructure to provide rapid integrated histomolecular diagnostics of gliomas. It is hoped that this infrastructure will allow for the introduction of clinical trials to test targeted therapies that are aimed at oncogenic 'driver' mutations and pathways of an individual's tumour (moving towards the concept of 'precision' or 'personalised' medicine for glioma patients).

I am the TJ-BM lead Biomedical Scientist and am responsible for sample receipt to the Oxford TJ-BM laboratory and preparing the samples for downstream applications such as histology, DNA and RNA extraction and Biobanking. Within this role, I have the opportunity to study for a higher degree in molecular neuropathology of CNS tumours. My project will expand on the scope of the TJ-BM project to assess the feasibility of analytical pathways for long read sequencing in neuropathology.

In the past, brain tumour classification has relied on microscopic features; this is now complemented by integration of molecular genetic features. Specifically, the field has been revolutionised by the development of an unbiased classification system that relies on the genome-wide methylation pattern of a tumour (Capper *et al*, 2018a). Currently, these tests can only be performed in a few centres and rely on expensive infrastructure. Turnaround time is slow. Developments in third-generation long read sequencing such as the PromethION sequencing device by Oxford Nanopore Technologies (ONT) can sequence DNA, including epigenetic modifications such as DNA methylation, in near 'real-time'. The aim is to perform long-read sequencing on a set of CNS tumours and compare and integrate the results with complementary data derived from the same samples. The hypothesis is that long read sequencing will result in novel insights into tumour biology and improved diagnostics, including reduced time to integrated histomolecular diagnosis.

1.2 Evolution of tissue diagnostics for tumours of the nervous system

Historically, tumours of any organ have been classified and diagnosed based on their (presumed) anatomical origin (e.g. a specific organ) and microscopic features (the anatomic-pathological approach). The early classification systems built on the work of Bailey and Cushing's classification system (Cushing and Bailey, 1927), which named tumours after the cell type in the developing embryo/foetus or adult which the tumour cell most resembled histologically (Collins, 2004). With this classification system, they aimed to understand if the microscopic features of gliomas carried different clinical implications. Following this, there were several attempts to establish an internationally accepted classification system that proved unsuccessful. However, in 1956 the World Health Organisation (WHO) executive board initiated the development of a uniform classification system that was to be used worldwide (Scheithauer, 2009).

The first edition of the WHO classification of Central Nervous System (CNS) tumours 'the blue book', published in 1979, was a collaborative effort lead by Professor Zulch and took almost a decade to complete (Zulch, 1979). This edition was based on the histological typing of tumours using light microscopic features in Haematoxylin and Eosin (H&E) and other simple stains. H&E is one of the most widely used stains and demonstrates the morphological phenotype of the lesion. While most tumours could be identified by conventional histological methods, a considerable number relied on other techniques such as Electron Microscopy (EM) (Coakham, Garson & Kemshead, 1984) EM allowed pathologists to study tumours based on ultrastructural features. This technique dominated the field until the late 1980's when it was substituted by immunohistochemistry (IHC). Thus,

the second edition of the ‘blue book’ published in 1993 incorporates the introduction of IHC.

Light microscopy IHC was described in 1941 (Coons, Creech & Jones, 1941) when the first fluorescent antibody labels were developed. In the late 1980’s the discovery of antigenic markers that are expressed on tumours (e.g. neurofilament) of the CNS lead to the advancement in monoclonal antibody technology (Coons, Creech & Jones, 1941); (Kohler & Milstein, 1975) that allowed for this technique to be become routine in diagnostic laboratories. The anatomic-pathological approach continued to form the basis of CNS tumour diagnostics. However, this approach has limitations and is insufficient for the age of precision or personalised medicine, which is based on the hypothesis that, in addition to anatomic-pathological features, it is the molecular signature of a tumour that may define its type and response to novel treatments (Malone *et al*, 2020).

As researchers began to understand the molecular origins of tumours and signatures that define subtypes and response to novel treatments (Malone *et al*, 2020), molecular markers began to be introduced into the classification of CNS tumours. One example of this was the discovery of the loss of chromosomal arms 1p and 19q as a molecular genetic alternation in one sub-type of glioma, oligodendrogliomas (Reifenberger *et al*, 1994). It was known that the majority of (morphologically defined) oligodendrogliomas responded well to a combination of chemotherapy (procarbazine, lomustine and vincristine (PCV)) (Cairncross & Macdonald, 1988). However, there were no clinical or histological markers that were able to predict with certainty this response to chemotherapy. Soon after this discovery, it was shown that the 1p/19q co-deletion signature in a glioma determines the sensitivity to PCV therapy and improved the over-all survival of patients (Cairncross *et al,* 1998). Following

this, molecular testing for this co-deletion was frequently used to classify this sub-type of glioma. However, even after the introduction of molecular makers, both the third and fourth editions of the blue book were still based on histological features.

Following the discovery of the structure of DNA in 1953, the ability to sequence DNA did not follow for some time. Many techniques were described in the 1970's but it was in 1977 that Sanger developed the chain-termination or dideoxy method to determine the sequence of nucleic acids (Sanger, Nicklen & Coulson, 1977). Sanger sequencing or first-generation sequencing was regarded as the gold standard for many years. However, the major drawback of Sanger sequencing was that it was extremely expensive to sequence a human genome. Early sequencing devices could only produce read lengths of between 250-500bp; however, newer platforms could generate longer read lengths of between 700-1000bp (Adams, 2008). It was estimated in 2006 the cost to sequence the human genome was approximately $14 million but the National Human Genome Research Institute (NHGRI) aimed to reduce the cost to $1000 (Schloss, 2008) and introduce technology to sequence the human genome faster. It was this initiative that led to second-generation or next generation sequencing (NGS).

NGS has allowed whole genome sequencing to become more feasible in the clinical settings due to the reduced cost and decrease in turnaround time which has been achieved by the ability to sequence many DNA fragments in parallel (Schwarze, Buchanan, Taylor & Wordsworth, 2018) (Schwarze *et al,* 2019). This technology has many applications including Whole-Genome Sequencing (WGS), Whole-Exome Sequencing (WES) or targeted panel sequencing (Schwarze *et al*, 2019). Like Sanger sequencing, the DNA strands are fragmented into short-reads lengths, however the read lengths are generally shorter (150-

300bp) than those for Sanger sequencing platforms. Short-read sequencing field has been dominated by Illumina technology by which DNA strands are sequenced by synthesis (Figure 1.1).

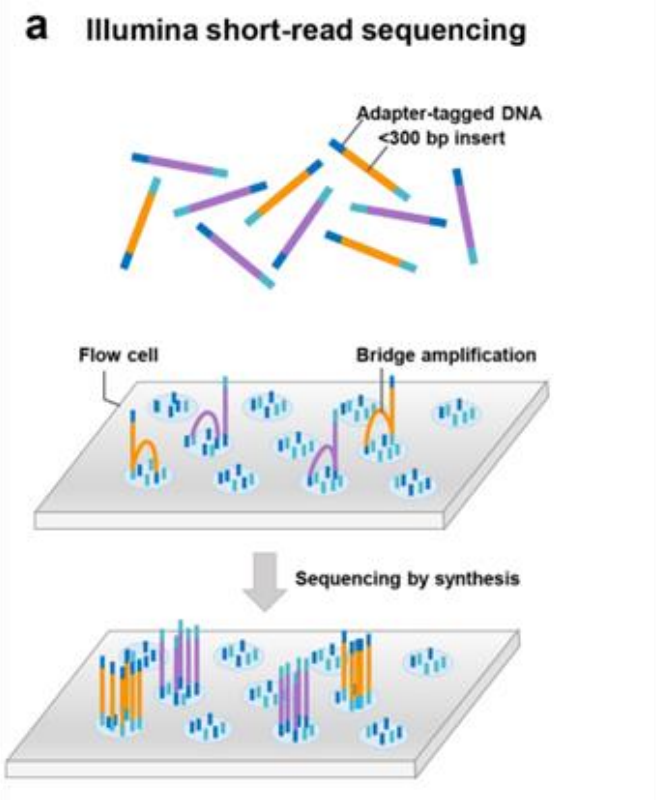


Figure 1.1 Overview of short-read sequencing technologies. DNA fragments are ligated to adapters that contain unique molecular identifiers and sequences complementary to the oligonucleotides that are attached to the surface of a flow cell. When the modified DNA is loaded onto a flow cell, the adapters ligated to the DNA fragments hybridize to the oligonucleotides on the flow cell surface. Copies of the DNA fragments are made by a process called bridge amplification by which the adapters on one end of a DNA fragment folds over and hybridizes to another, oligonucleotide in the flow cell. A polymerase adds nucleotides to build double-stranded bridges of DNA fragments. This is repeated, generating millions of clusters of double stranded DNA. Following this the reverse strands of DNA are cleaved and washed away leaving only forward strands. Sequencing by synthesis begins by a fluorescently labelled deoxyribonucleotide triphosphates (dnTNPs) are incorporated into the newly synthesize DNA stations. A laser excites the fluorophore on the strand which emits a characteristic fluorescent emission signal that corresponds to the base (figure adapted from Logsdon, Vollger, Eichler, 2020

The increasing ease of applying genomic technologies to tumour samples has resulted in an explosion of molecular genetic data, which, in general terms, have taught us that previously anatomico-pathologically defined tumour 'entities' may in fact represent several distinct tumour types. It is due to this development in knowledge and technologies that the concept of an integrated histomolecular diagnosis.

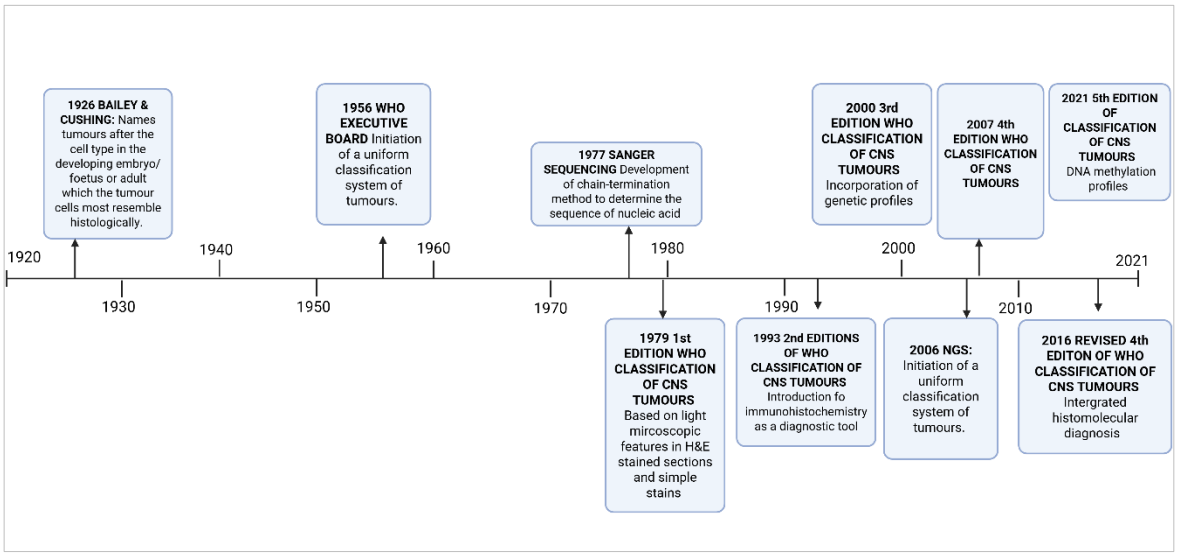


Figure 1.2 Schematic diagram to illustrate how tumour diagnostics has evolved over the last 100 years.

1.3 The concept of an integrated histomolecular diagnosis

The concept of a so-called integrated histomolecular diagnosis has been embraced in the 2016 WHO classification of tumours of the nervous system, the first iteration of the WHO classification mandating the identification of specific molecular markers to define a tumour type (e.g. IDH mutation and 1p/19q codeletion for the diagnosis of oligodendroglioma) (Louis *et al,* 2016).

The International Society of Neuropathology suggested that the diagnosis should be a four-tiered layered approach (Figure 1.3) including integrated diagnosis (layer 1), histological classification (layer 2), WHO grade (layer 3) and molecular information (layer 4)(Louis *et* al, 2014). As there is no limit on the molecular data that can be included in layer 4, and it allows for discrepancies between layer 2 or 3 and the integrated diagnosis, this approach allows flexibility.

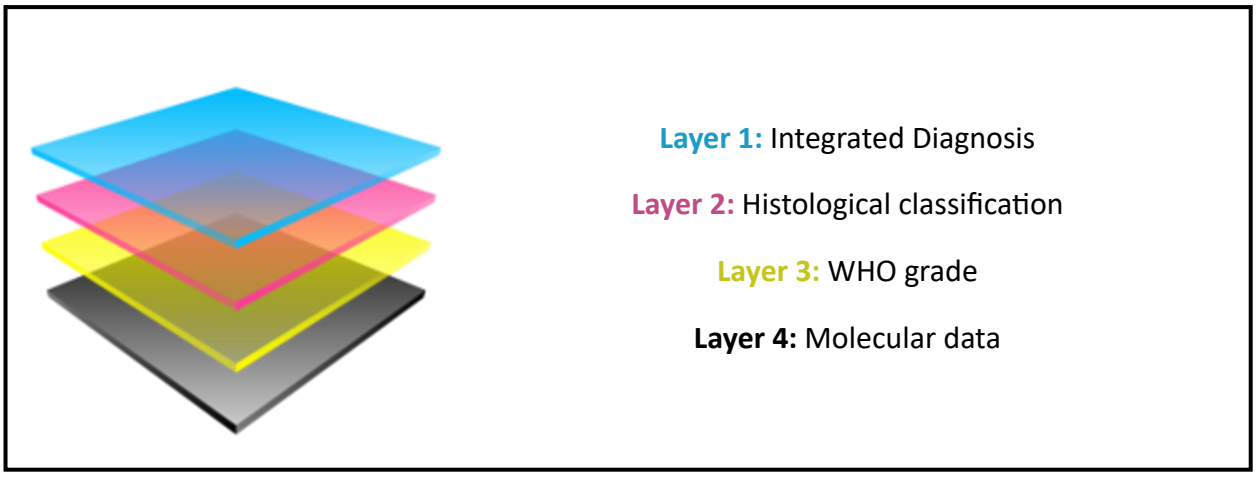


Figure 1.3 Diagram to show the concept of the integrated histomolecular diagnosis. A histological diagnosis, WHO grade and molecular information are integrated to form layer 1, integrated diagnosis.

The most recent (2021) edition of the WHO classification of CNS tumours incorporates DNA methylation profiles for those tumour entities for which the methylome offers diagnostic guidance (Louis *et al,* 2021). Methylation profiles can be used to diagnose majority of brain tumours and can be used to characterise rare tumour types that are difficult to diagnose by morphological and molecular features alone. Using methylation profiles has helped identify new tumour entities and subclasses that were previously not recognised. For example, the fifth edition of WHO classification of CNS now recognises four subtypes of medulloblastoma which can be defined molecularly (WNT-activated, Sonic Hedgehog (SHH)- activated, group 3 and group 4) (Ellison and Taylor, 2021) in addition to four subtypes (classic, desmoplastic nodular, extensive nodularity and anaplastic) defined histologically (Ellison and Taylor, 2021; Torp, Solheim and Skjulsvik, 2022).

1.4 Epigenetic modifications – DNA methylation

Epigenetic modifications are alterations to genes without altering the DNA. The concept of epigenetic modifications was first described by Waddington in 1942 (Waddington, 2012). He described the importance of both genetics and the environment in determining cell fate during embryonal development (Figure 1.4). DNA methylation was the first epigenetic modification to be described (Holliday and Pugh, 1975). DNA methylation is the addition of a methyl group (CH_3) to the cytosine-guanine dinucleotides (CpG) (Newell-Price, Clark and King, 2000. DNA methylation is essential for normal development as methylation patterns can modify gene expression and cell differentiation. However, alterations to DNA methylation patterns can result in dysregulation to cellular processes which in turn can lead to disease (Newell-Price, Clark and King, 2000; Moore, Le and Fan, 2013). CpG islands are regions of DNA with higher than usual frequency of CpG repeats which are usually unmethylated in normal cells and are often found within the promotor regions of genes. In cancer cells, CpG islands often become hypermethylated with the addition of a methyl group to the 5^{th} carbon of a cytosine to form 5-methylcytosine (5mC) which leads to silencing of transcription factors and inactivation of gene promotors (Esteller, 2000). Oxidation of 5mC produces 5 hydroxymethylcytosine (5hmC), another form of DNA methylation. Unlike 5mC, which is associated with gene silencing, 5hmC is linked to DNA demethylation and active gene expression (Greenberg and Bourch'his, 2019; Colquitt *et al*, 2013; He, Yao and Yi, 2024).

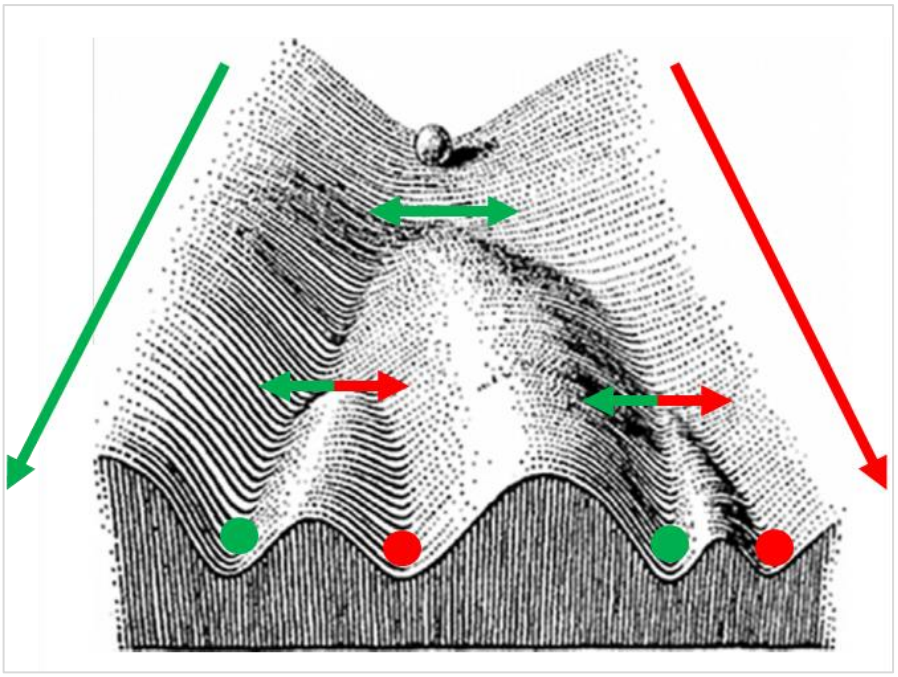


Figure 1.4 The epigenetic landscape as described by Waddington (1957). Cell differentiation can be described by the metaphor of a cell traveling down a riverbed. The cell which is represented by a ball can take different trajectories or cell fates. I have modified Waddington's original drawing with the aim to illustrate that upon neoplastic transformation (red arrow/ball) of a cell during its normal evolutionary trajectory (i.e. rolling down the hill of differentiation, green arrow/ball) it retains epigenetic signatures of its differentiation state and acquires new epigenetic marks reflecting the neoplastic state. This explains in a simplified way why the brain cancer methylome reflects signatures of the cell of origin, including topography, and marks of neoplasia induced by pathogenic genomic variants.

The llumina HumanMethylation450 BeadChip (450K) was the most widely used array for assessing DNA methylation in human samples. This BeadChip initially covered 450,000 CpG sites but was later upgraded to the MethylationEPIC BeadChip which uses the same technology as the 450k but covers 850,000 CpG sites including over 90% of the sites on the 450k chip. However, the 450k chip primarily covered promoter regions whereas the EPIC chip also covers gene enhancer regions (Pidsley *et al,* 2016). This technology is based on sodium bisulphite conversion of DNA in which an unmethylated cytosine is converted into uracil, following PCR amplification, uracil is detected as thymine during sequencing. Methylated cytosines (5mC) are not converted and therefore remain as cytosines during sequencing which allows them to be distinguished from unmethylated cytosines (Li and Tollefsbol, 2011).

In clinical settings, formalin fixed paraffin embedded (FFPE) tissue blocks are more widely accessible than fresh frozen tissue and can be used as the input material for the EPIC chip. DNA from FPPE is known to be of poor quality due to the process of fixing tissue in formalin causing DNA protein cross-links to preserve the morphology of the tissue. This means that the library preparation needs to include a step to repair the DNA either before or after bisulphite conversion (de Ruijter *et al*, 2015).

Whilst the EPIC array is a valuable diagnostic tool, it does have limitations. The array is limited to the 850k CpG sites where as long-read sequencing platforms such as Oxford Nanopore Technologies (ONT) can detect methylation changes across the whole genome.

Methylation classification needs to be reproducible in clinical pathology settings. A brain tumour methylation classifier was developed at the German Cancer Research Center (DKFZ) and Heidelberg University (Capper *et al*, 2018a; Capper *et al,* 2018b). The Heidelberg classifier is a machine learning approach that uses random forest algorithms which are trained binary decision trees. This classifier recognised 91 classes in version 11 (v11) and expanded to 184 subclasses in version 12.8 (v12.8) and is currently available through a free online tool (Capper *et al*, 2018a; Benfatto *et al*, 2025; Sill *et al*, 2026). This allows for a comprehensive and reproducible approach for DNA methylation classification of all CNS tumour entities.

MGMT gene (O6-methylguanine-DNA methyltransferase) methylation is found in many cancers including glioma. It is located on chromosome 10 (10q26) and methylation of MGMT promoter and enhancer region is a molecular prognostic marker of gliomas. Epigenetic modifications, more specifically of CpG islands within the MGMT promotor regions is linked to silencing of the gene (Hermes *et al,* 2008; Yu, Zhang, Wei and Shao,

2020). MGMT gene codes for a repair enzyme that transfers a methyl group from guanine to cysteine which avoids cell death and tumorigenesis caused by alkylating agents (Yu, Zhang, Wei and Shao, 2020). MGMT methylation has been found to be a prognostic marker and predictive of the response to alkylating chemotherapy agent, temozolomide (Esteller *et al*¸2000*;* Hegi *et al,* 2005; *Leske et al*, 2023). Methylation of the promotor region is typically found in patients with long-term survival when compared to patients with short-term survival rates. In addition, studies have shown that patients with MGMT methylation show better response to temozolomide treatment due to the methylation silencing the gene which ultimately leads to improved survival (Malley *et al*, 2011; Leske *et al,* 2023).

In addition to epigenetic brain tumour classification, DNA epigenetic classifiers have also been developed for other tumour types such as sarcomas (Koelsche *et al*, 2021). Similarly to the Heidelberg classification of brain tumours, the sarcoma classifier is a machine learning algorithm built on methylation data from 62 tumour methylation classes. Unlikely the brain tumour classifier that is routinely used to aid diagnosis, the sarcoma classifier has limitations including misclassifications, and so it is not currently widely used for clinical diagnosis (Roohani *et al,* 2022).

DNA methylation classification is incorporated into the most recent version of WHO classification of CNS tumours. However, methylation profiling is performed separately to whole genome sequencing and other molecular tests. Long-read sequencing has the potential to incorporate methylation classification and whole genome sequencing into one clinical test.

1.5 Pre- analytical phase in neuropathology

The Pre-analytical phase in neuropathology encompasses all steps prior to analytical testing, including specimen collection, transportation and tissue processing. These steps are critical in determining the quality and reliability of downstream analyses. Optimal fixation is essential to preserve cellular architecture, enabling accurate histological assessment of the tumour, while simultaneously maintaining nucleic acid integrity for molecular techniques such as whole genome sequencing. Current tissue preservation approaches include fixation in 10% neutral buffered formalin, which preserves cell morphology by an alcohol fixation phase followed by cross-linking of proteins (Thavarajah *et al,* 2012), and snap freezing, often using liquid nitrogen and isopentane, to maintain molecular integrity. The integrity of nucleic acids preserved during this phase directly impacts DNA extraction quality, which is particularly important for advanced molecular techniques such as long-read sequencing technologies that require high molecular weight DNA. Standardisation of these procedures is therefore essential to ensure that tissue specimens are suitable for both morphological assessment and molecular investigation.

1.6 The potential of long-read sequencing for diagnostics and research

The most recent developments in genetic sequencing, third-generation or long-read sequencing technology offer novel biological insights as well as practical innovations. This field is currently led by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). Unlike Sanger and NGS where short-read lengths make it almost impossible to accurately assemble short reads from regions of high guanine-cytosine content or regions of large structural variants (SV), long read sequencing technologies can overcome these limitations. As the name suggests, long-read sequencing platforms have the advantage of producing long reads of greater than 10kb (Van Dijk, Jaszczyszyn, Naquin & Thermes, 2018) (Figure 1.5). Also, as long-read library preparations and sequencing do not require PCR amplification, base modifications such as methylation can be detected as the DNA/RNA is in its original state (Schadt, Turner & Kasarskis, 2010; Mantere, Kersten, & Hoischen, 2019).

Long read sequencing has the potential to be used in clinical settings to identify molecular markers that are otherwise difficult to resolve by conventional sequencing techniques (Figure 1.5). Large structural variants are difficult to detect using short-read sequencing techniques as the reads are too short to cover the entire length of the variant and therefore the breakpoints are often missed leading to inaccurate detection, however, length of the of reads for long-read platforms, allows for these variants to be detected more accurately (Romagnoli, Bartalucci and Vannucchi, 2023; Moustakli *et al*, 2025). Similarly, repeat expansions including areas repetitive CG regions are difficult to detect by short-read as again the read lengths will not cover the repeat regions. Long-read has the ability to sequence across these repetitive regions.

Haplotype phasing is the process in which variants can be mapped to the chromosome of origin to determine which variants are located on the same chromosome (Buckley, Ideker, Carter and Schork, 2019). Long-read sequencing has improved phasing and is another improvement of long-read approaches.

In addition to structural variant, repeat expansion and phasing detections, long-read has the capability to detect pseudogenes. These are DNA regions that resemble functional genes but are in fact an inactive form due to mutation within the sequence (Lynch and Conery, 2000). Long-read sequencing platforms can sequence over the full gene transcript and can therefore distinguish the pseudogene from the parent genes (Troskie *et al,* 2021).

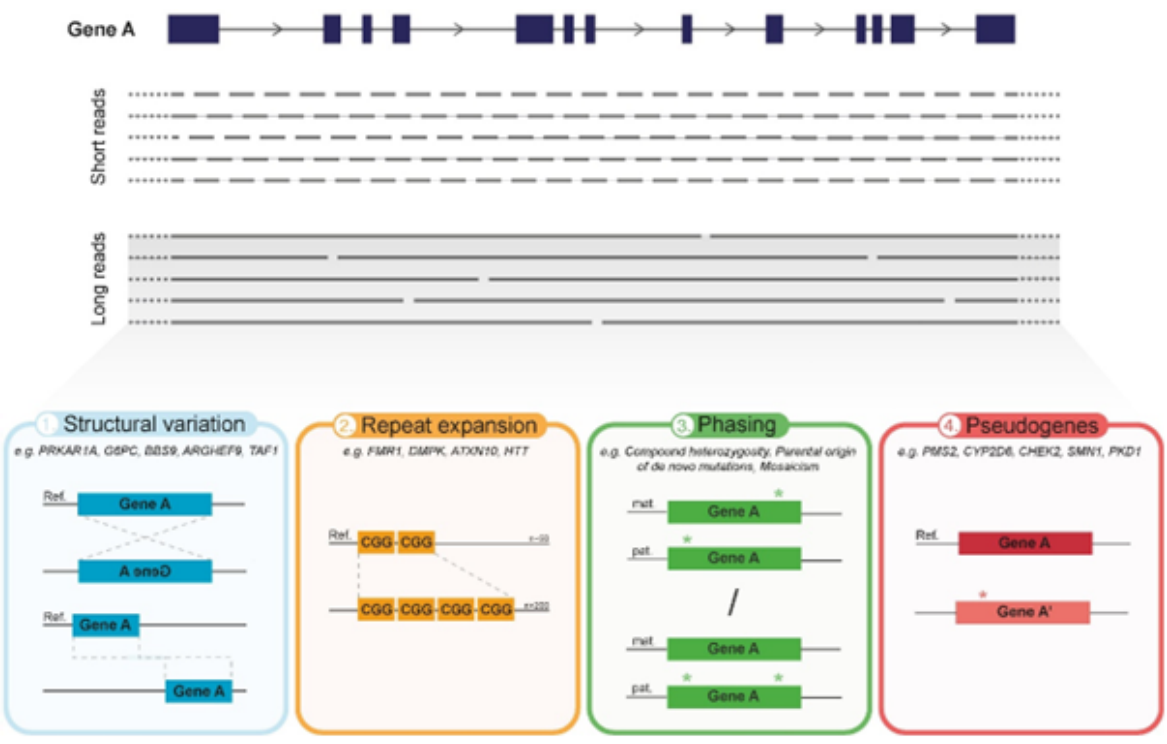


Figure 1.5 Overview of the advantages of long-read sequencing approaches. The main difference between long and short read sequencing is the increase in read length. Long-read sequencing devices have the ability to sequence reads over 10kb in length in compared to short-read platforms (150-330bp). **1.** Improved detection of large structural variations due to longer read lengths and no need for PCR. **2.** Capacity to sequence over long tandem repeat expansions and extreme GC-rich regions. **3.** Enhanced phasing by assignment of genetic variants to the homologues paternal or maternal chromosomes, to determine patterns of inheritance, parental origin of de novo events, mosaicism, allele specific expression and disease

risk haplotypes. **4.** Identify clinically relevant genes from their pseudogenes. (Adapted from Mantere, Kersten, & Hoischen, 2019).

Even though long-read sequencing platforms have advantages compared to short-read sequencing platforms, the platforms have faced challenges, the main being the high error rates when compared to short-read sequencing platforms. Illumina short-read platforms have an error rate of <1% whereas in the early releases of both PacBio and ONT, the error rate was 10-20% (Zhang, Jain and Aluru, 2020). Updates in the ONT nanopore designs (flow cell versions from R9.4.1 to R10.4.1) and improvement in basecalling software, the error rate has improved to <10% (Santos *et al*, 2025).

1.7 Oxford Nanopore Sequencing

The concept behind ONT was described in 1996 when researchers at Harvard showed translocation of DNA or RNA molecules through a nanopore protein channel (Brandin, Branton, and Deamer, 1996; Deamer, Akeson and Branton, 2016). Deamer and Branton recognised that α-hemolysin which is a toxin released by the bacterium Staphylococcus aureus could form membrane channels that had a diameter that would allow only one strand of DNA to pass through. After substantial work on α-hemolysin, it was deduced that due to the shape of this molecule, the individual bases could not be accurately identified (Meller, Nivon and Branton, 2001). It was concluded that further improvements to this method could result in detection of nucleotide sequences in single molecules of RNA or DNA relatively quickly.

The first sequencing device using nanopore technology was released commercially in 2015 was the small compact device, MinION. Since then, three larger devices (GridION, PromethION, and PromethION 48) have been released (Mantere, Kersten, & Hoischen, 2019).

Table 1.1 Shows specifications of ONT devices. A flongle uses an adapter to attach to the MinION device but uses a flongle flow cell which allows for low depth cost efficient sequencing. The MinION is a compact device that plugs into a laptop or pc with the main benefit being a portable device that can be used in a lab or field setting. GridION is a benchtop device that can use up to 5 MinION flow cells. PromethION line of devices are bench top devices capable of using multiple flow cells. The p2 device needs to be used in parallel with a PC whereas the P24 and P48 have inbuilt computer systems.

Device name	Number of flow cells per device	Number of channels / pores	Yield (Gb) per flow cell
Flongle	1	126 channels	~3 Gb
MinION	1	512 channels / 2048 pores	Up to 50 Gb
GridION	Up to 5	512 channels / 2048 pores	Up to 50 Gb
PromethION 2 (P2)	Up to 2	2675 channels / 12,000 pores	100 – 200 Gb
PromethION 24 (p24)	Up to 24		
PromethION 48 (p48)	Up to 48		

All ONT devices use flow cells containing nanopores approximately one nanometre in diameter, allowing single strands of DNA to pass through. Nanopores are protein structures which are embedded into an artificial membrane with an electrolytic solution (Wang *et al,* 2021). When a current is applied to the membrane, the DNA strand is able to pass through the nanopore which causes a disruption to the current. Each of the four nucleotides bases are different size and shapes and therefore cause different disruptions to the current and a unique pattern. These patterns can then be decoded by basecalling algorithms, and the DNA sequence can be determined. Epigenetic modifications to DNA such as the addition of a methyl group which cause a distinct disruption to the current and therefore can also be detected (Figure 1.6).

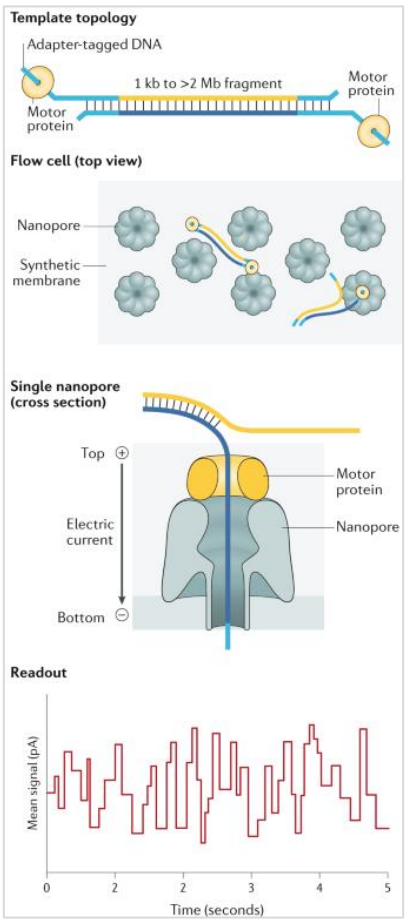


Figure 1.6 How ONT sequencing platforms work. The electrical current that flows through each individual nanopore can be measured. When DNA is translocated through a nanopore it disrupts the current. Each nucleotide base causes a different electrochemical characteristic which allows each nucleotide to be distinguished from one another, allowing identification of the DNA sequence. This process is in real time which allows the users to end the run when enough data has been obtained (Adapted from Logsdon, Vollger and Eichler, 2020).

ONT offer a range of library preparation kits that vary in in-put material (DNA, RNA, cDNA), library preparation time and throughput. The kits use one of two types of reagents: ligation or transposase to attach adapters to the double stranded DNA. The rapid sequencing kit uses transposes-based reagents to apply the adapters and the fragmentation of the DNA in the same step which enables for quick library preparation which can be completed within 10 minutes. Whereas the ligation kit requires two distinct steps, first the fragmentation of the DNA followed by ligation of the adapters with a library prep time of approximately 60 minutes. Both kits do not require a PCR step which allows for base modifications such as methylation to be detected during sequencing. Barcoding kits are available that include

unique DNA barcode adapter sequences ligated to each sample, allowing multiple samples to be pooled and sequenced together.

1.8 Single-nuclei sequencing

Bulk whole genome sequencing and epigenomic profiling provides important insights into the molecular landscapes of CNS tumours by producing sequencing data that is representative across the tumour sample. Tumours are extremely heterogeneous meaning they contain diverse cell populations which each have different genetic and epigenetic traits (Ren, Kang and Zhang, 2018). This issue can be resolved by single cell sequencing which is able to detect the gene expression of an individual cell. CNS tumour samples are difficult to dissociate due to the complex tumour architecture. In addition, archived samples are normally frozen which causes rupture of cell membranes leading to cell death and reduced cell viability (Slyper *et al.* 2020; Zhou *et al*, 2024). As a results, these tissue types are often unsuitable for single cell studies. An alternative to single cell sequencing is single nuclei sequencing which allows the analysis of genetic traits of individual nuclei rather than intact cells.

Single nuclei workflows consist of four main stages, the first being isolation of the nuclei. For this process, there must be a balance between successfully isolating the nuclei but also preserving the nuclear integrity (Kersey *et al,* 2025). Isolation is a combination of a mechanical process by grinding of the tissue and the use of lysis buffers to ensure the tissue is fully dissociated. The lysate is then filtered to remove cellular debris prior to nuclei counting. The second stage is barcoding of the individual nuclei. 10x Genomics (10X Genomics, USA) use a droplet-based method which uses barcoded gel beads that are encapsulated into droplets with the single nuclei forming Gel Beads-in-Emulsion (GEMs) on a chip as shown in Figure 1.7 (10x Genomics, Chromium GEM-X Single Cell 5’ Reagent Kits v3, 2024). Once the individual nuclei are barcoded, nuclear RNA is reverse transcribed into

cDNA and amplified to ensure sufficient material for library preparation. ONT have released a single-cell/nuclei sequencing protocol in conjunction with 10x Genomics which allows for library construction using the ONT ligation sequencing kit (Oxford Nanopore Technologies, Single-cell transcriptomics sequencing from 5' cDNA prepared with 10x Genomics using SQK-LSK114, 2025). ONT has key advantages over short-read approaches most importantly being long-read allows for full length transcripts to be sequenced. These libraries can be sequenced at a high depth on a PromethION flow cell which will generate long-reads with the individual nuclei barcodes and unique molecular identifiers (UMI's). A UMI is a unique tag which is added to each individual cDNA molecule during the library prep procedure before PCR amplification. The addition of the individual barcodes and UMI allows for accurate quantification of the genomic features of a single nucleus (Zheng *et al*, 2017).

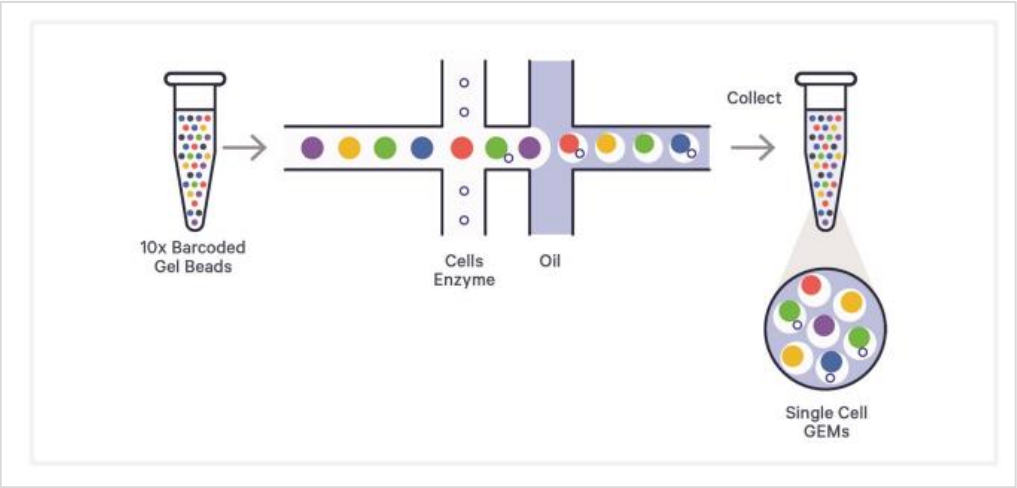


Figure 1.7 10x Genomics GEM formation. Adapted from Chromium GEM-X Single Cell 5' Reagent Kits v3 User Guide *(10x Genomics, Chromium GEM-X Single Cell 5' Reagent Kits v3, 2024).*

The final stage for single-nuclei sequencing is data analysis. The sequencing data is processed through bioinformatic pipelines including alignment to a reference genome, quality control, normalization and clustering to identify distinct cell types and the genomic

traits of single nuclei or cell populations. Figure 1.8 shows an overview of the 10x Genomics and ONT single nuclei workflow.

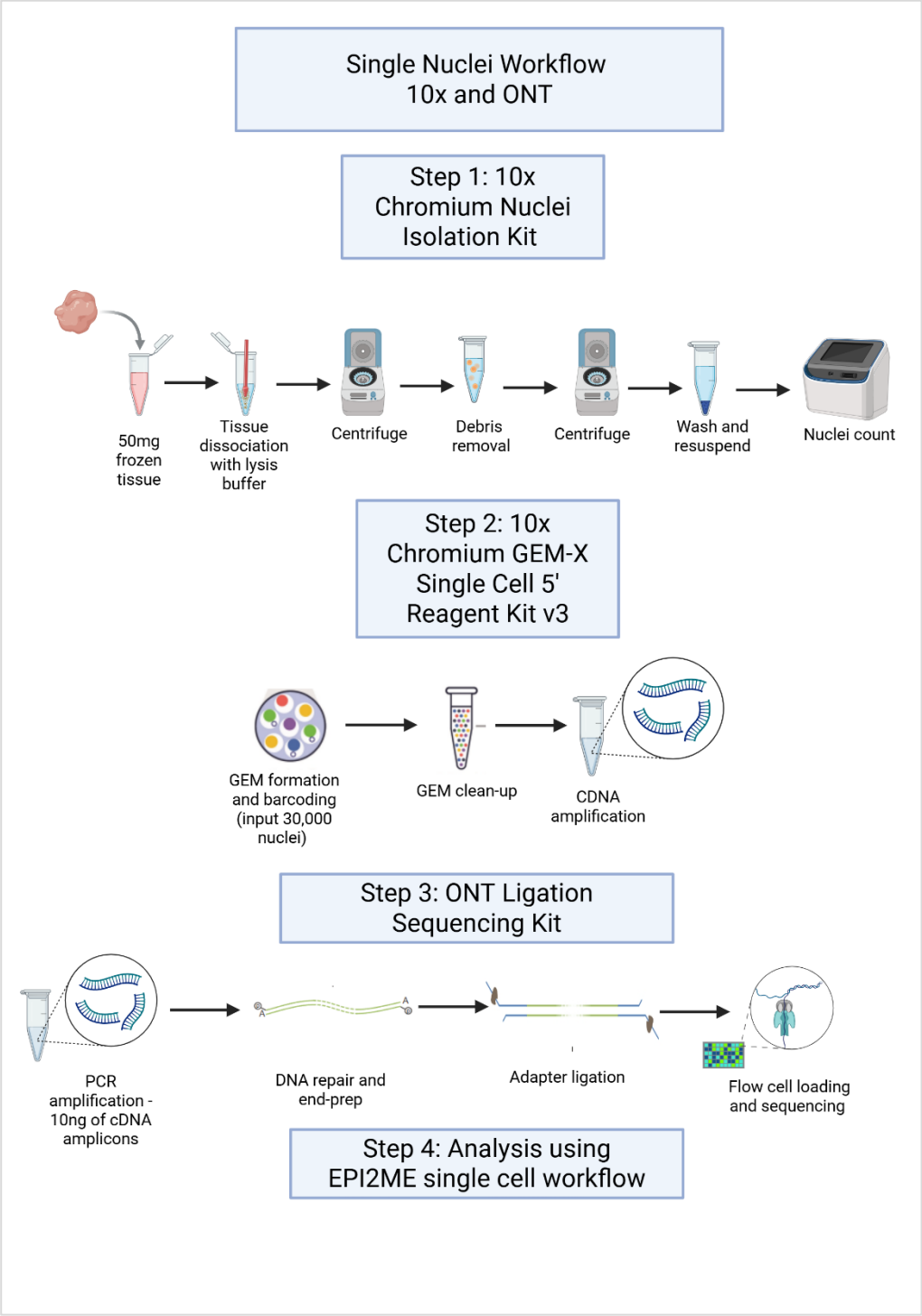


Figure 1.8 Single-nuclei RNA sequencing workflow combining 10x Genomics and Oxford Nanopore Technologies (ONT). *Frozen tissue is processed using the 10x Genomics Nuclei Isolation kit to release nuclei through lysis followed by debris removal, washing, resuspension and nuclei counting. Purified nuclei are then encapsulated using 10x Chromium GEM-X Single Cell 5' Reagent Kit v3 where nuclei undergo GEM formation, barcoding , clean-up and cDNA amplification. Amplified cDNA is prepared for long-read*

sequencing using ONT ligation sequencing kit. Sequencing data is analysed using EPI2ME single-cell workflow. Figure created with BioRender.com.

1.9 Tumours of the central nervous system

Tumours of the CNS are a diverse group of neoplasms that arise in the brain, spinal cord or the surrounding structures. These tumours can be primary neoplasms which originate in the CNS or metastatic lesions in which the primary tumour originates elsewhere, and tumour cells migrate to CNS tissues forming metastases. Compared to all cancer types, primary tumours are the CNS are uncommon and only account for 3% of cancer case annually in the UK (Wanis *et al*, 2021). Tumours of the CNS can be grouped by the cell of origin such as glioma, meningioma and embryonal tumours Louis *et al,* 2016; Louis *et al*, 2021; Torp, Solheim and Skjulsvik, 2022). Treatment of CNS tumours usually involves a multi-model approach including surgery, radiotherapy and chemotherapy. However, as previously mentioned, these tumours are extremely difficult to treat due to the location of the tumours and the blood brain barrier reducing the efficacy of chemotherapy drugs that are successful treatments of cancers of other origins (Bender, 2018; Han *et al,* 2020).

1.9.1 Glioma

Glioma are the most common primary tumours of the central nervous system that originate from glial cells in the brain and spinal cord (Weller *et al,* 2024). They predominantly affect adults between 45-70 years and are more predominant in males (Molinaro *et al*, 2019). Gliomas are diagnosed based of their histological features and molecular signatures (Louis *et al,* 2021; Sahm *et al*, 2023). The 2021 WHO classification of central nervous system tumours (Louis *et al,* 2021; Weller *et al,* 2024) further divide glioma into six subgroups; Adult-type diffuse gliomas, Paediatric-type diffuse low-grade gliomas, Paediatric-type diffuse high-grade gliomas, Circumscribed astrocytic gliomas; Glioneuronal and neuronal tumours, and Ependymal tumours.

Adult-type diffuse gliomas include Astrocytoma, IDH-mutant, Oligodendroglioma, IDH-mutant and 1p/19q-codeleted, and Glioblastoma, IDH- wildtype. The molecular markers are important as they can be used as predictors of prognosis, for example, IDH mutated gliomas are generally associated with better prognosis where as IDH-wildtype glioblastomas tend to progress more rapidly and have poor survival (Hu *et al,* 2025).

1.9.2 Embryonal tumours

Embryonal tumours are a group of highly malignant CNS neoplasms that arise from primitive, undifferentiated neuroepithelial cells in the brain and spinal cord (Louis *et al,* 2021; Weller *et al,* 2024). These tumours predominantly affect children and adolescents. The 2021 WHO classification of central nervous system tumours (Louis *et al,* 2021; Weller *et al,* 2024) classifies embryonal tumours into several subtypes; medulloblastoma (which can be further subdivided by molecular features); Atypical teratoid/rhabdoid tumour (ATRT), Cribriform neuroepithelial tumour, Embryonal tumour with multilayered rosettes, CNS neuroblastoma - FOXR2-activated and other rare entities. Among these groups, medulloblastoma is the most common subtype. Medulloblastoma are divided into four molecular subgroups; WNT-activated, SHH-activated, Group 3, and Group 4 (Northcott *et al,* 2012; Louis *et al,* 2021).

1.9.3 Mesenchymal Chondrosarcoma

Mesenchymal Chondrosarcoma (MCS) is rare tumour that affects children and young adults with no targeted treatments. MSC have an incidence rate of 0.2-0.7 cases per 100,000 and comprises only 1% of all chondrosarcoma cases (Shakked *et al,*2012; Hunter *et al,* 2016). There have been less than ten cases diagnosed in Oxford University Hospitals Trust since 1991. These tumours have a poor prognosis with a 5

year survival rate of 70%, 10-year survival rate of <55%, however, the presence of metastases at diagnosis significantly affect prognosis and median survival falls from 20 years for localised disease to 3 years for metastatic disease (Frezza *et al,* 2015).

MCS have a widespread anatomical distribution but mostly occur in craniofacial or spinal region (Fanburgh-Smith, Ladanyi, de Pinieux, 2020; Kleinschmidt-DeMasters *et al,* 2021). The WHO classification of CNS tumours histologically describes MCS tumours as being characterised by the presence of a biphasic microscopic architecture consisting of undifferentiated small blue cells and areas of differentiated hyaline cartilage as shown in Figure 1.9 (Fanburgh-Smith, Ladanyi, de Pinieux, 2020; Kleinschmidt-DeMasters *et al,* 2021). To distinguish this tumour type from others that have similar pathology of small blue cells, immunohistochemistry is used for the detection of SOX9 protein, a positive stain is a characteristic of MCS (Wehrli *et* al, 2003; Fanburgh-Smith, Ladanyi, de Pinieux, 2020; Kleinschmidt-DeMasters *et al,* 2021).

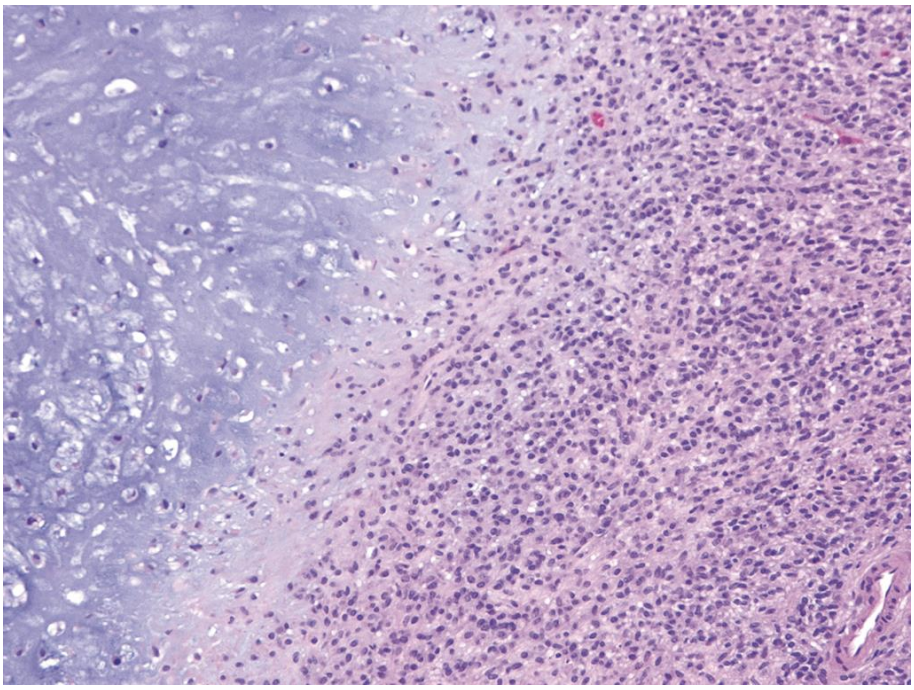


Figure 1.9 Classical biphasic histology of MCS with spontaneous transition from poorly differentiated mitotically active areas (right) to better differentiated less proliferative areas (left) which may even acquire mature cartilage-like morphology (inset). *Reproduced from Kleinschmidt-DeMasters et al, 2021, WHO Classification of Tumours: Central Nervous System Tumours.*

Wang *et al*, (2012) identified the fusion of two genes; *HEY1* and *NCOA2* (Figure 1.10) as characteristic molecular marker of this tumour type. Both genes are located on chromosome 8 (8q21.13 and 8q13.3) and the fusion is caused by a 10 Mb deletion resulting in a fusion at exon 4 of *HEY1* and exon 13 of *NCOA2* (Wang *et al,* 2012; Nakayama *et al,* 2012). The fusion gene occurs in >80% of MCS cases (Wang *et al,* 2012; Nakayama *et al,* 2012; El Beaino *et al,* 2018; Xu *et al,* 2022) and is listed as a desirable criterion for the diagnosis of MCS but it is not disease defining. However, this may become disease defining in the future as larger MCS cohorts are investigated for molecular characteristics. Currently, the largest cohort to molecularly confirm MCS was published by Xu *et al,* 2022 and included molecular study of 13 cases.

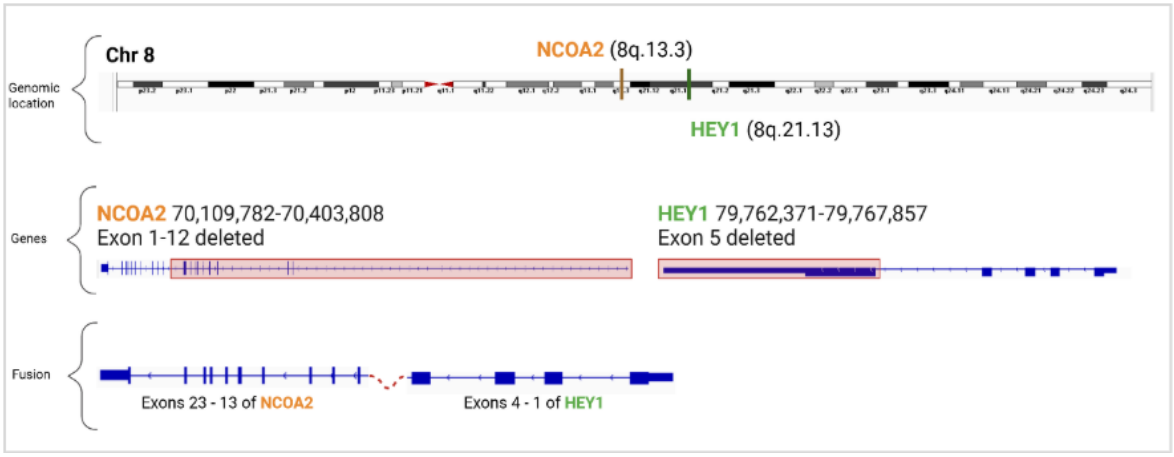


Figure 1.10 ***HEY1::NCOA2 fusion in MCS.*** *Both genes are located on chromosome 8 which leads to an intrachromosomal deletion of exon 1-12 of NCOA2 and exon 5 of HEY1.*

Reverse Transcriptase-Polymerase Chain Reaction (RT-PCR) and Fluorescent In situ Hybridization (FISH) were the conventional methods used to detect *HEY1::NCOA2* gene fusion. RT-PCR is a process that converts RNA to cDNA and then uses PCR primers that are designed to span the specific fusion to amplify the region. FISH detects fusions by using custom probes that are labelled with fluorescent markers which bind to markers within the fusion. More recently, the Archer FusionPlex panel (Seager, Aisner, and Davies 2020), which is a solid tumour panel that uses NGS of RNA for the detection of chromosomal rearrangements and gene fusions has been used for fusion detection in solid tumours. Fusion detection is enabled by using a gene specific primer that target the exon of a gene of interest. One advantage of this panel over the RT-PCR or FISH technique is that as the panel does not require the use of primers for both fusion partners, only one of the gene partners needs to be known allowing detection of gene fusions without prior knowledge of the fusion partner or the exact breakpoint. There are over 100 genes on this panel and test can be order on subset of genes or a single gene.

As the molecular architecture of these tumour types is still unknown, there is a lack of therapeutic treatment options for patients. Primary tumours are treated by surgical

resection of the lesion, and in some cases, combined chemotherapy and radiotherapy may be considered. However, there are no established guidelines for patients' metastatic stage of the disease (Dudzisz-Sledz *et al,* 2023).

1.10 Aims of thesis

Tissue preservation and quality control (Chapter 2)

1) To identify and test novel ways of preanalytical tissue handling optimised for rapid ONT diagnostics and research.

Real time diagnostics and monitoring of brain tumours (Chapter 3)

1) To develop infrastructure in Oxford to provide rapid integrated histomolecular diagnosis in hours rather than weeks.

Molecular pathological architecture of MCS (Chapter 4)

1) To understand the molecular pathological architecture of MCS using bulk and single-cell long read sequencing.

Chapter 2 Tissue preservation and quality control

2.1 Aims of chapter

Classic biobanking of both surgical biopsies and postmortem tissue relies on snap freezing tissue at collection. However, this approach may lead to freezing-induced rupture of outer cell membranes and loss of viability of cells. PrestoCHILL is a freezing technique that can prevent the formation of ice crystal artefact. Furthermore, shipment of frozen tissue on dry ice is both complex and expensive. The latest development in tissue-specific, commercially available transport and culture media may allow the shipment of tissue samples at room temperature if restricted to transit of 24-48 hours. Thus, one aim of this project is to identify and test novel ways of preanalytical tissue handling optimised for long-read sequencing for rapid ONT diagnostics and research.

Our hypothesis is that use of novel culture media and freezing techniques will result in superior preservation of tissue and molecules compared to existing methods such as freezing in liquid nitrogen vapour or formalin fixation.

2.2 Introduction

For long-read sequencing to be successful, the optimal sample preparation including tissue preservation and DNA extraction needs to be determined. Long-read sequencing relies on the extraction high quality, high molecular weight DNA. The requirements for ONT sequencing method are described in Table 2.1.

Table 2.1. gDNA requirements for ONT sequencing protocols.

Sequencing protocol	DNA input requirements	260/280 ratio	Flow cells loading requirements	Loading volume	Molecular weight
Ligation Sequencing kit (WGS)	**2 ug gDNA**	**1.8 – 2.0**	**300 ng**	**32 µl**	**>10kb**
Ligation Sequencing kit (Single-cell)	**10 ng cDNA**	**N/A**	**50 – 100 fmol or 33 ng of library**	**32 µl**	**1Kb**
ROBIN Ultralong sequencing kit (adapted protocol)	**300 ng gDNA**	**1.8 – 2.0**	**All available library**	**45 µl**	**>10kb**

The first challenge we face is to be able to preserve the tissue in a way that protects the integrity of the DNA to meet the requirements of downstream analysis by long-read technologies but also allows the tissue to be suitable for other downstream analysis such as single cell-type transcriptomic studies. This is crucial for the future of tissue diagnostics as pathologists aims to employ the full range of novel 'omics' technologies, beyond genomic analysis. Currently, the gold-standard method to preserve tissue (either surgical biopsies or post-mortem tissue) for genetic analysis is freezing the samples in pre-cooled isopentane or liquid nitrogen vapour. This method of preservation allows for extraction of high molecular weight DNA and DNA yields that are suitable for whole genome sequencing (Grizzle, Bell & Sexton, 2011). However, along with the health and safety risks incurred by

this freezing method, this approach may result in freezing-induced rupture of outer cell membranes and loss of viability of cells which limits what downstream applications the tissue is suitable for.

PrestoCHILL (Milestone^{TM}, Italy) is a cryoembedding system (figure 2.1) that allows freezing of tissue in under 60 seconds and could improve tissue quality by preventing the formation of ice crystal artefact. The speed in which the tissue is frozen means that tissue samples are immediately available for downstream analysis such as ONT sequencing.

The tissue is placed directly onto a metal surface that is pre-chilled to -40^{o}c which allows rapid freezing but also provides the user with a flat surface which provides cryoblocks that are easily cut, unlike conventional techniques which require trimming of the tissue after freezing to provide a flat surface for cutting.

Figure 2.1 Image of PrestoCHILL instrument. *The tissue sample is placed onto aluminium molds for rapid freezing (approximately 60 seconds). The instrument features a touch screen interface for setting tissue processing parameters.*

For the comparison of sequencing techniques part of this trial, tissue samples would need to be shipped to Oxford from multiple centres around the UK. Frozen tissue can be transported overnight in dry ice. However, this method of transportation is costly. Therefore, we aim to test if the latest generation of culture media for neural tissue and cells are suitable for transport of fresh brain tissue obtained at surgery or post-mortem using postal shipping at ambient temperature.

The second preanalytical challenge that is faced, is the ability to extract long high molecular weight DNA. Many commercially available kits for DNA extraction cause fragmentation on DNA strands which are suitable for short read sequencing but not long read but also long-read sequencing platforms are sensitive to contaminants bound to DNA following extraction using kits which can cause shorter read sequences. We aim to compare two commercially available extraction kits; Monarch® HMW DNA Extraction Kit for Tissue (New England Biolabs®, UK) and Zymo Quick-DNA Microprep Kit (Zymo Research, USA).

In this chapter, the effect of tissue preservation and DNA extraction techniques was only tested on DNA, not RNA, as DNA was the required input material for long-read whole genome sequencing used in chapters 3 and 4. Additionally, the extraction methods tested (Monarch HMW DNA Extraction Kit and Zymo Quick-DNA Microprep Kit) are designed to isolate DNA exclusively and do not recover RNA. As such, this chapter focuses specifically on evaluating how difference preservation strategies impact DNA integrity and yield.

Table 2.2 DNA extraction kit specifications for Monarch HMW DNA extraction kit for tissue and Zymo Quick DNA microprep kit

	Standard input	**Average yield**	**Average DNA length**	**Cost per kit**
Monarch® HMW DNA Extraction Kit for Tissue	10-25 mg	5-20 µg	50-≥500 kb	£484.00 (50 preps)
Zymo Quick-DNA Microprep Kit	5 mg	5 µg	Up to 40kb	£145.00 (50 preps)

2.3 Methodology

2.3.1. Cohort

Post- mortem tissue samples used in this study were donated to The Oxford Brain Bank, a research tissue bank within the University of Oxford. All patients gave informed consent for tissue to be used for medical research, including genetic analysis. This part of the study was conducted under The Oxford Brain Banks generic ethical approval (Oxford C REC 23/SC/0241). Fresh human brains are received by the brain bank prior to processing of the tissue into frozen and fixed blocks.

Surgical tissue biopsies used in this study were surplus to diagnostic requirement. These samples were accessed under ethical approval from HRA and Health and Care Research Wales (24/WS/0013).

2.3.2 Tissue preservation in culture media

Tissue samples were obtained from human post-mortem brain tissue (n=8) and surgical biopsy (n=1) to assess if the latest generation of culture media for neural tissue is suitable for transport of fresh brain. Post-mortem tissue was used as it allows access to large amounts of material. All tissue was processed fresh at the time of donation. The specific brain regions differed depending on the clinical diagnosis as described in Table 2.3.

Table 2.3 Case selection for samples used in tissue preservation trial.

Case	Clinical diagnosis	PM delay	Brain region	Preservation medium	Freezing method
NP003-2021	Dementia	31hrs	Cerebellum	Hibernate-A for 24/96hrs at RT or 4°c	Liquid nitrogen vapour and isopentane
NP004-2021	Glioblastoma, IDH Wild-type, WHO grade IV	38hrs	A2, tumour infiltration zone	Hibernate-A for 24/96hrs at RT or 4°c	Liquid nitrogen vapour and isopentane
NP008-2021	Glioblastoma, Giant cell variant, IDH wild-type, WHO grade IV	52hrs	P3, tumour and infiltration zone, mesyl temporal zone	Hibernate-A for 24/96hrs at RT or 4°c	Liquid nitrogen vapour and isopentane

NP013-2021	Control	13hrs	A3, frontal cortex	Hibernate-A for 24/96hrs at RT or 4°c	Liquid nitrogen vapour and isopentane
NP024-2021	MND	28hrs	M1/S1	Atelerix Tissue Ready for 24/96hrs at RT	Isopentane
NP029-2021	MND	32hrs	M1/S1	Atelerix Tissue Ready for 24/96hrs at RT	Isopentane
NP038-2021	Glioblastoma	70hrs	A4, tumour nodule	Hibernate-A for 24/96hrs at RT or 4°c and Atelerix Tissue Ready for 24/96hrs at RT	Isopentane
SH1049-2025	Astrocytoma	N/A (surgical sample)	Tumour	Atelerix Tissue Ready for 24/96hrs at RT 72hrs	Slow freezing

Fresh tissue samples were dissected into fragments of 0.5 x 0.5 x 0.5 cm and were stored in Gibco™ Hibernate™-A Medium (Gibco; Fisher Scientific, USA) and/or TissueReady™ (Atelerix, UK). Hibernate-A culture medium was supplemented with 2% (v/v) B27 (Gibco, ThermoFisher Scientific). TissueReady™ was prepared following the manufactures protocol (Atelerix TissueReady™ Handbook, March 2020). This protocol required the gel base beads to be diluted 1:5 (v/v) with a cell culture medium. Hibernate-A supplemented with B27 (2% v/v) was used. The PM tissue samples were stored in the preservation media for either 24 or 96 hours. The tissue workflow for samples stored in sample media is shown in Figure 2.2. Samples in Hibernate-A were stored at either room temperature (RT) or 4°c and samples in Atelerix TissueReady at RT as indicated in the manufacture protocol. From each PM case, immediately after fresh dissection, one tissue fragment was placed directly into Neutral buffered formalin solution (4% w/v formaldehyde) (Genta) and one frozen in liquid nitrogen vapour to act as our control as these are the current tissue preservation methods. Following storage of the tissue in the preservation media, samples were either fixed in Neutral Buffered Formalin and processed into formalin fixed paraffin embedded (FFPE) blocks or snap frozen by one of two methods: Liquid nitrogen vapour, or pre-cooled isopentane.

DNA was extracted from 30mg of PM tissue using the AllPrep DNA/RNA Mini Kit (Qiagen, Germany) following the manufactures protocol. DNA was eluted in 50µl EB buffer. DNA concentration was measured using the Qubit 4 Fluorometer (ThermoFisher Scientific, USA). DNA samples were prepared using the Qubit™ dsDNA BR Assay Kit. The purity of DNA was measured using the Nanodrop 2000/2000c Spectrometer (ThermoFisher Scientific, USA). DNA fragment size was measured using 2100 Bioanalyzer (Agilent Technologies, USA). Samples were prepared using Agilent Technology High Sensitivity DNA kit following manufacturing protocols.

For surgical biopsy tissue, the sample was dissected into four fragments of 0.5 x 0.5 x 0.5 cm, and 3 fragments were stored TissueReady™ (Atelerix, UK). TissueReady™ was prepared as described above. The tissue samples were stored in the preservation media for 72 hours. 1 fragment was slow frozen at -80°C immediately after fresh dissection. Following storage of the tissue in the preservation media, samples were either fixed in Neutral Buffered Formalin and processed into formalin fixed paraffin embedded (FFPE) blocks, slow frozen at -80°C or prepared for DNA extraction. DNA was extracted from 100 mg of tissue that had been stored in TissueReady and 100 mg of tissue that was slow frozen after fresh dissection using Monarch® HMW DNA Extraction Kit for Tissue (New England Biolabs®, UK) following the manufactures protocol. DNA was eluted in 100 µl Monarch Elution buffer II (New England Biolabs®, UK). DNA concentration was measured using the Qubit 4 Fluorometer (ThermoFisher Scientific, USA). DNA samples were prepared using the QubitTM dsDNA HS Assay Kit. The purity of DNA was measured using the Nanodrop 2000/2000c Spectrometer (ThermoFisher Scientific, USA). DNA fragment size was measured using Agilent Tapestation 4200 (Agilent, USA). Samples were prepared using Genomic DNA regents and screentape (Agilent, USA) following manufactures protocols.

To assess the quality of tissue PM and surgical tissue, 6µm section from FFPE blocks were stained with Haematoxylin and Eosin (H&E) using an automated Shandon Linistain GLX Random Access Stainer (ThermoFisher Scientific, USA).

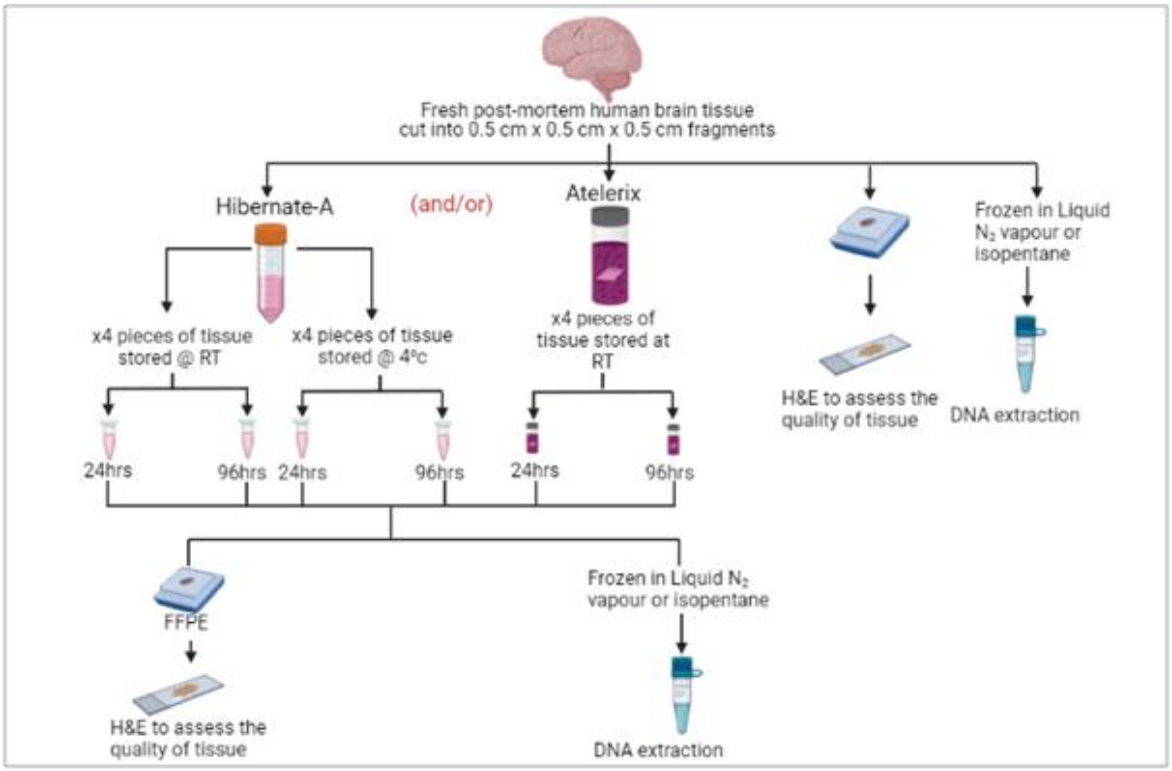


Figure 2.2 Tissue workflow of head-to head comparison of the latest generation of culture media for neural tissue for transport of fresh brain tissue. *Tissue samples were stored in either Hibernate-A (Gibco, Fisher Scientific) or TissueReady (Aterlerix) for 24 hrs or 96 hrs at RT or 4°C. Following storage, samples were either snap frozen for DNA extraction or fixed in formaldehyde for H&E assessment of the quality of tissue. Figure created with BioRender.com.*

2.3.3 Tissue freezing using PrestoCHILL

Tissue samples were obtained from human post-mortem brain tissue samples (n=14), human biopsy (n=9) and mouse tissue samples (n=9) to assess to quality of tissue frozen by PrestoCHILL compared to standard freezing methods. All tissue was processed fresh at the time of retrieval.

Fresh tissue samples were dissected into fragments of 1 x 1 x 1 cm and were frozen at -40°c for 60 seconds on the PrestoCHILL cryoembedding system. Tissue samples either had no cryo preservation medium, a small amount of Milestone Cryoembedding Compound (MCC) ((Mileston™, Italy), a small amount of CellPath OCT Embedding Matrix (Fisher Scientific, UK) or embedded fully in MCC or OCT (conditions shown in Table 2.4). To assess the quality of tissue, 10 µm sections from the frozen blocks were cut immediately after freezing and

stained with H&E using a Leica Autostainer XL (Leica Biosystems, UK). Following this, frozen tissue blocks were stored in a -80^{o}c for 1 week before sectioning to reassess the quality of the tissue. At the initial time of freezing, a tissue fragment of the same size was placed into liquid nitrogen vapour, the standard way of freezing to act as a control. Figure 2.3 shows the tissue workflow.

Three surgical cases (SH894/2025, SH912/2025 and SH946/2025) had DNA extracted using QIAamp Fast DNA Tissue Kit (Qiagen, Germany) following the manufactures protocol. DNA from these samples were used for real time monitoring of brain tumour using ONT sequencing (Chapter 3).

Table 2.4 Cryopreservation conditions when freezing tissue with PrestoCHILL.

Case	Tissue type	Diagnosis	Brain region	Cryopreservation medium (MCC or OCT)	Freezing method
NP049-2025	**Post-Mortem**	**Parkinson's Disease**	Cerebellum	None	PrestoCHILL
			Cerebellum	MCC on cryo cork	PrestoCHILL
			M1/S1	MCC – embedded	PrestoCHILL
			M1/S1	MCC on cryo cork	PrestoCHILL
			Cerebellum	MCC – embedded	PrestoCHILL
			Cerebellum	OCT – embedded	PrestoCHILL
			Cerebellum	None	Liquid N_2 vapour
NP050-2025	**Post-Mortem**	**Alzheimer's Disease**	Basal ganglia	None	PrestoCHILL
			Cerebellum	None	PrestoCHILL
			Cerebellum	None	PrestoCHILL
			M1	None	PrestoCHILL
			M1	None	PrestoCHILL
			Basal Ganglia	None	PrestoCHILL
			Cerebellum	None	Liquid N_2 vapour
SH894-2025	**Surgical biopsy**	**Ependymoma**	Tumour	MCC – embedded	PrestoCHILL
			Tumour	None – tissue tube	PrestoCHILL
SH912-2025	**Surgical**	**Astrocytoma**	Tumour	None	PrestoCHILL
			Tumour	None	PrestoCHILL
			Tumour	None	PrestoCHILL
SH946-2025	**Surgical**	**Low grade glioma**	Tumour	None	PrestoCHILL
			Tumour	None	PrestoCHILL
SH947-2025	**Surgical**	**Oligodendroglioma**	Solid tumour	None	PrestoCHILL
			CUSER tumour	None	PrestoCHILL
			Sagittal	MCC – embedded	PrestoCHILL
			Sagittal	MCC – embedded	PrestoCHILL

Mouse 1	**Mouse**	**N/A**	Sagittal	MCC – embedded	PrestoCHILL
Mouse 2	**Mouse**	**N/A**	Muscle	None	PrestoCHILL
			Muscle	MCC – embedded	PrestoCHILL
			Sagittal	MCC – embedded	PrestoCHILL
			Sagittal	None	PrestoCHILL
			Sagittal	None	PrestoCHILL
			Sagittal	None	PrestoCHILL

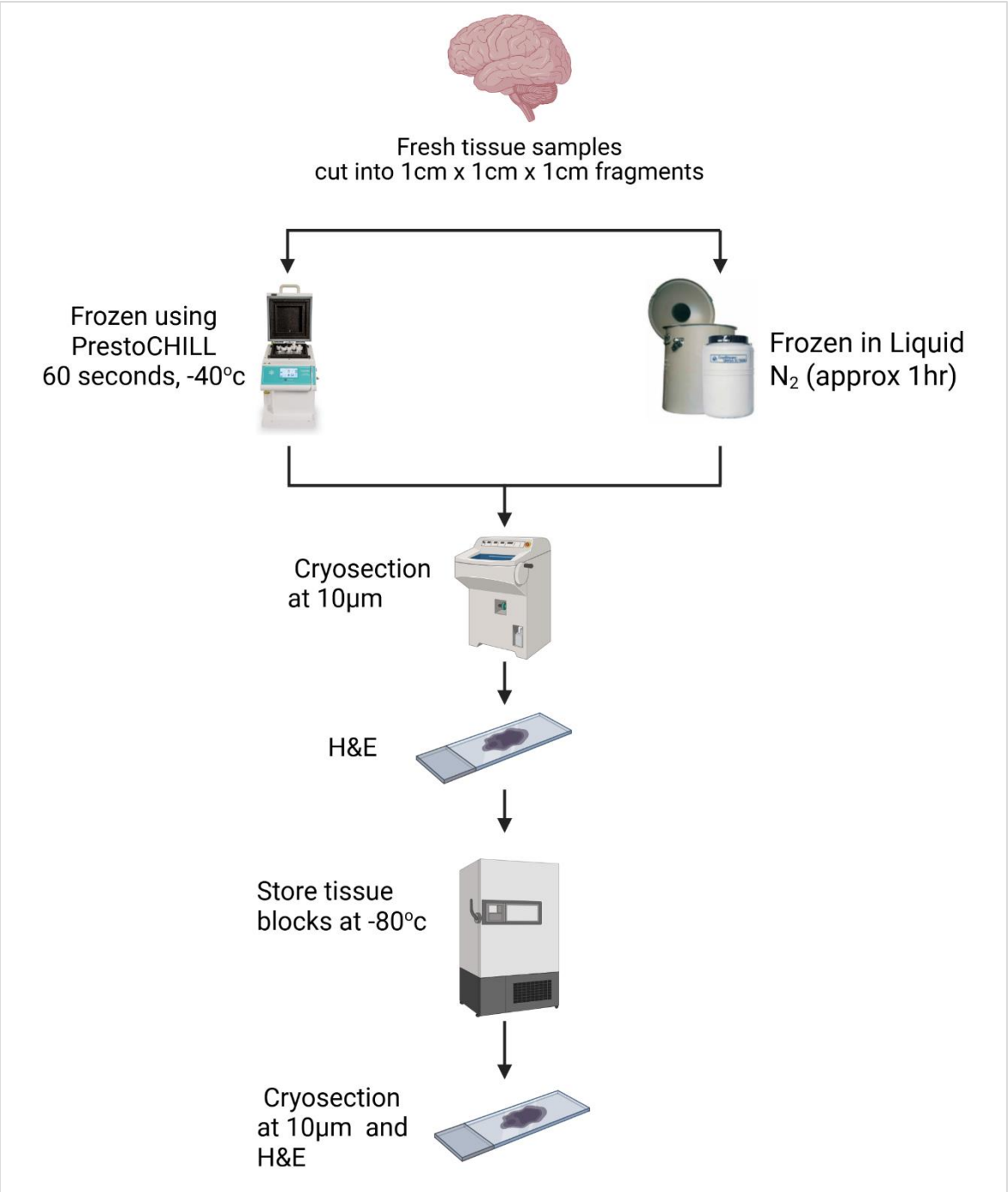


Figure 2.3 Tissue workflow of head-to head comparison freezing techniques. *Tissue samples frozen either by PrestoCHILL (Milestone) or in liquid nitrogen vapour. Samples were sectioned for H&E assessment of the quality of tissue. Following this, samples were stored at -80°c before tissue quality was assessed to see how storage at -80°c could affect the quality of the tissue. Figure created with BioRender.com.*

2.3.4 DNA extraction using Monarch® HMW DNA Extraction Kit for Tissue and Zymo Quick-DNA Microprep Kit

To test DNA extraction kits, tissue samples were obtained from frozen surgical biopsies of confirmed Mesenchymal Chondrosarcoma (MSC) (n=8). Mesenchymal Chondrosarcoma samples were received from University of Leiden, Department of Pathology. These cases were used under ethical approval from London – Stanmore Research Ethics Committee (REC 17786; IRAS 45163).

DNA was extracted from x40 10 µm sections of tissue using Monarch® HMW DNA Extraction Kit for Tissue (New England Biolabs®, UK) following the manufactures protocol. DNA was eluted in 100 µl Monarch Elution buffer II (New England Biolabs®, UK). DNA was extracted from x40 10 µm sections of tissue using Zymo Quick-DNA Microprep Kit (Zymo Research, USA) following manufactures protocol. DNA was eluted in 15 µl DNA elution buffer (Zymo Research, USA).

DNA concentration was measured using Invitrogen Qubit 4 Fluorometer (ThermoFisher Scientific, USA). DNA samples were prepared using the Invitrogen Qubit™ dsDNA HS Assay Kit (ThermoFisher, Scientific). The purity of DNA was measured using the Nanodrop 2000/2000c spectrometer (ThermoFisher Scientific, USA). DNA fragment size was measured using Agilent Tapestation 4200 (Agilent, USA). Samples were prepared using Genomic DNA regents and screentape (Agilent, USA) following manufactures protocols.

2.3.5 Statistical analysis

Statistical analysis was performed using GraphPad Prism statistical software version 10.6.0 (GraphPad Software). P values <0.05 were considered statistically significant. Absorbance

ratio (260/280) and DNA concentration data from tissue preservation workflow were analysed with a two-way ANOVA with the Geisser-Greenhouse correction. The analysis was performed to compare timeframes within the tissue medium (Control, 24hrs and 96hrs), storage temperatures of Hibernate-A (RT and 4°c) and comparing the two preservation media at RT (Hibernate-A and TissueReady). To compare time frames for samples stored in TissueReady, a one-way ANOVA with the Geisser-Greenhouse correction and paired t-test were used. To compare time and temperature variables of samples stored in Hibernate-A a one-way ANOVA with the Geisser-Greenhouse correction and paired t-tests were used.

PM Datasets with storage times other than 24 or 96hrs were excluded from the statistical analysis.

For the comparison of DNA extraction kits, normality was tested using D'Agostino Pearson omnibus normality test, a paired t-test was used to test significance of the two groups.

2.4 Results

2.4.1. Tissue preservation in culture media

Histological assessment of samples stained with H&E did show some variability between the two preservation media (Figure 2.4 A-F). There was no difference in tissue quality between tissue placed directly into formaldehyde and tissue preserved in Hibernate-A for 24 and 96hrs at RT. Samples stored in TissueReady was of poorer microscopic quality in our initial experiments on normal brain tissue (Figure 2.4 A-F). The difference concerned a larger extent of microvacuolation (Figure 2.4 E and F). Samples were placed in Hibernate-A medium immediately following dissection, however, as TissueReady had to be made up 5 minutes before use, samples were left out at room temperature before being placed into the TissueReady gel. This could have affected the quality of tissue as the samples dried out. I adapted our protocol to be able to place tissue directly into TissueReady gel following dissection; this resulted in the elimination of the presumed 'drying-out' microvacuolation artefact (Figure 2.4 H-I).

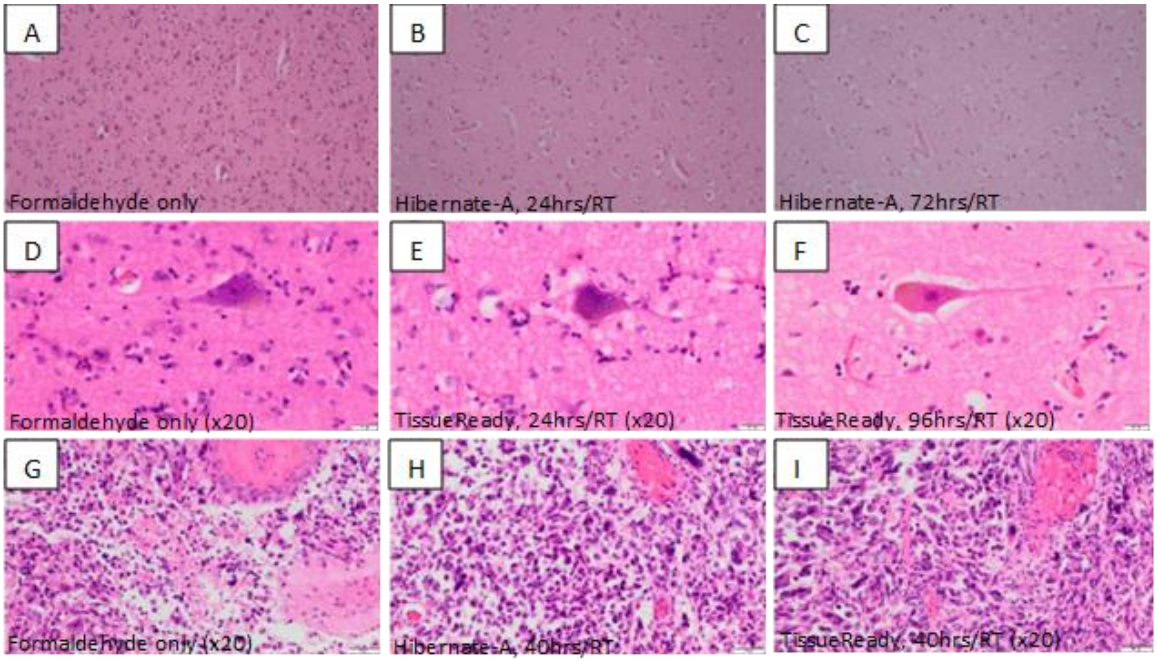


Figure 2.4 Microscopic images of H&E stained post-mortem brain tissue following storage in different preservation media. (Formaldehyde, Hibernate-A or TissueReady) *A-C: NP013/2021, tissue stored in Formaldehyde or Hibernate-A at RT for 24 or 72hrs; D_F: NP024/2021, tissue stored in Formaldehyde or TIssueReady at RT for 24 or 96hrs; G-I: NP038/2021, tissue stored in Formaldehyde, Hibernate-A or TissueReady at RT for 40hrs.*

The DNA purity was assessed using 260/280 ratio and DNA concentration. 260/280 ratio for DNA extract from NP013/2021 was excluded from statistical analysis as Nanodrop readings could not be obtained.

The 260/280 absorbance ratio for DNA extracted from tissue stored in Hibernate-A varied between 1.37 and 2.01 and tissue stored in TissueReady varied between 1.68 and 2.69 (Figure 2.5). There was no statistical difference between the two preservation media at different time points and temperatures. There was no statistically significant differences of 260/280 ratio between both time (control, 24hrs or 96hrs) or temperature (RT or 4°c) stored in Hibernate-A (two-way ANOVA with Geisser-Greenhouse correction, P value >0.05). When time (control vs 24hrs, Control vs 96hrs, 24hrs vs 96hrs) and temperature (RT and 4°c) were compared separately using a paired-t, the 260/280 absorbance ratio of tissue stored in Hibernate-A for 96hrs at RT was statistically significant compared to the control (P

value 0.0335). There was also no significant difference between the samples stored in TissueReady compared to samples frozen immediately after dissection (one-way ANOVA with Geisser-Greenhouse correction, and paired t-test, P value >0.05).

Concentration of DNA extracted from tissue stored in Hibernate-A decreased the longer tissue was stored in the medium (0hrs, 24hrs, 96hrs) for both temperature variations (RT and 4°c). This was not the case for DNA extracted from tissue stored in TissueReady. There were no statistically significant difference between the different temperatures (Hibernate-A RT, Hibernate-A 4°c, TissueReady RT) or between the time points (Control, 24hrs and

96hrs) two-way ANOVA with the Geisser-Greenhouse correction and paired t-test, P value >0.05) (Figure 2.6).

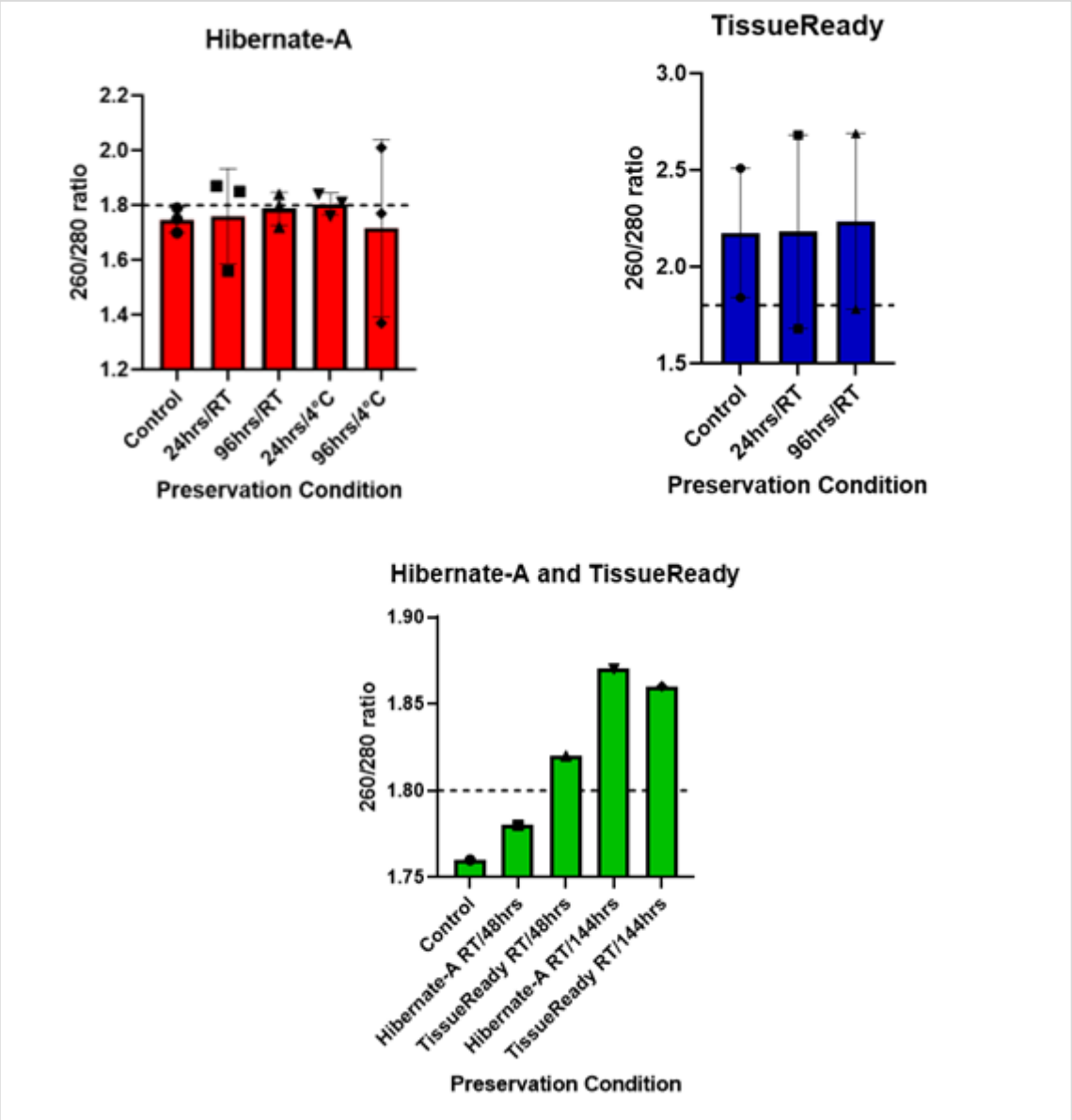


Figure 2.5 Data showing DNA quality by absorbance ratio at 260/280 from DNA extracted from tissue stored in different preservation media. *DNA was extracted from human brain tissue following storage in different preservation mediums (Hibernate-A or TissueReady) at RT 4°c for 24 or 96hrs. DNA was extracted using AllPrep DNA/RNA Mini Kit (Qiagen) and the 260/280 absorbance ratio measured using Nanodrop spectrophotometer. Hibernate-A n=3, TissueReady n=2, both Hibernate and TissueReady n=1. Error bars = SEM. Samples from once case stored in Hibernate-A were only stored for 72hrs not 96hrs. These results were excluded from the statistical analysis.*

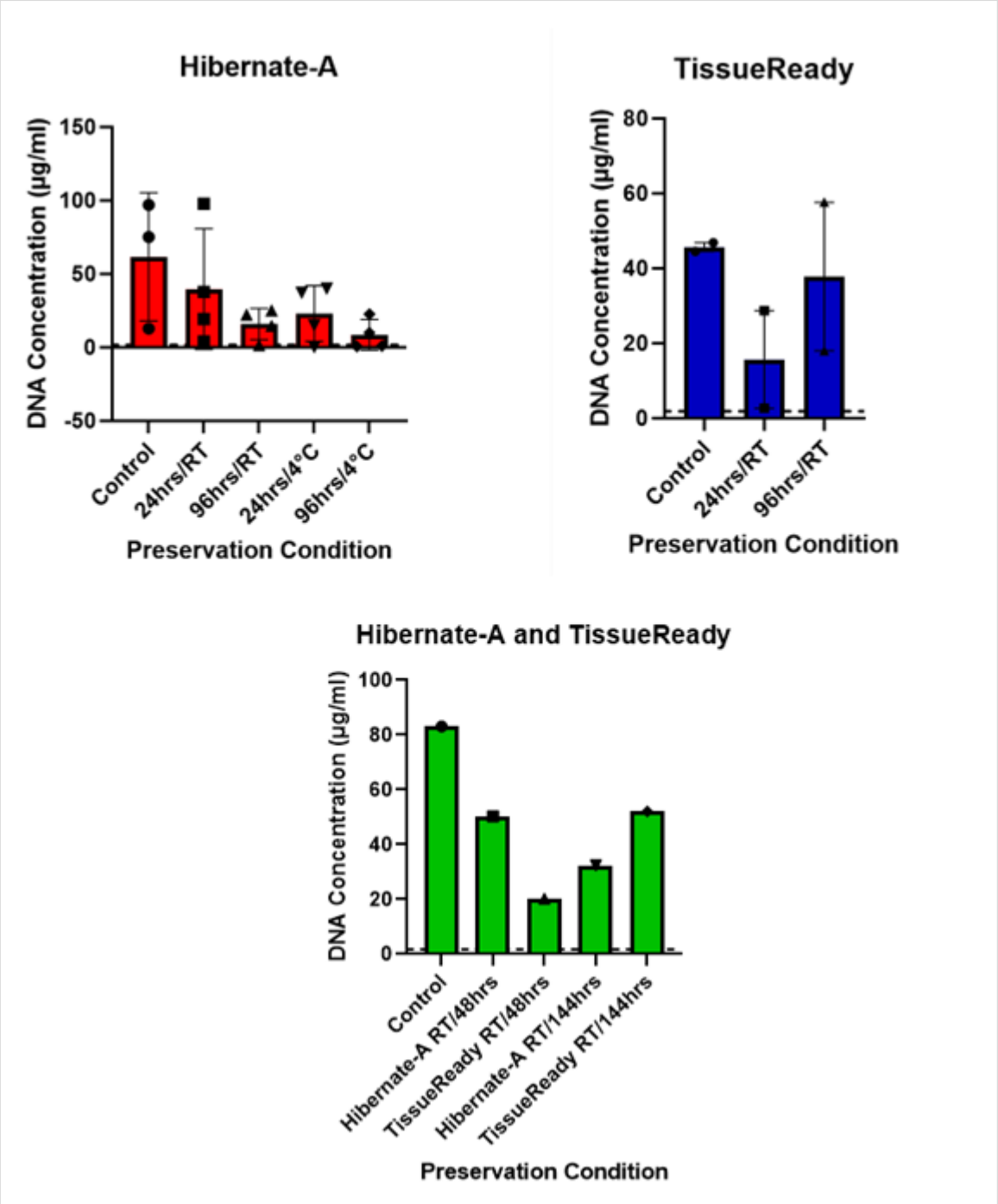


Figure 2.6 Data showing DNA concentration of DNA extracted from tissue stored in different preservation media. *DNA was extracted from human brain tissue following storage in different preservation mediums (Hibernate-A or TissueReady) at RT or 4°c for 24 or 96hrs. DNA was extracting using AllPrep DNA/RNA Mini Kit (Qiagen) and the DNA concentration measuring using Quibt 4 fluorometer. DNA samples were prepared using Qubit dsDNA BR Assay Kit. Hibernate-A n=2, TissueReady n=2, both Hibernate and TissueReady n=1. Error bars = SEM. Samples from one case stored in Hibernate-A were only stored for 72hrs not 96hrs. These results were excluded from the statistical analysis.*

2.4.2. Freezing using PrestoCHILL

All samples were successfully frozen using PrestoCHILL within 60 seconds which is significantly quicker than the time it takes to freeze tissue samples with liquid nitrogen vapour. Histological assessment of the frozen cryosections using H&E (Figure 2.8) showed ice crystal artefact was absent in 100% of the tissue samples and cell morphology was retained when frozen in 60 seconds using PrestoCHILL. When compared to tissue samples frozen with liquid nitrogen vapour, ice crystal artifact was observed (Figure 2.8). Following storage of the frozen tissue samples at -80°C for 1 week, there was no deterioration observed histologically and tissue architecture remained intact (figure).

When comparing tissue embedded using MCC, OCT or no tissue embedding medium, there was no difference in the quality of the tissue. A difference in the ease of tissue handling was noted. Samples without embedding medium or a small amount of MCC or OCT showed better stability and were easier to section on the cryostat. Tissue samples fully embedded into the medium showed poor adherence to the tissue and cracked post freezing making sectioning difficult (Figure 2.7).

DNA extracted from tissue samples frozen using PrestoCHILL had a 260/280 ratio between 1.83 and 1.84 which is considered pure DNA. The yield was between 3.3 µg and 7.4 µg which is sufficient DNA for input material for ONT whole genome sequencing.

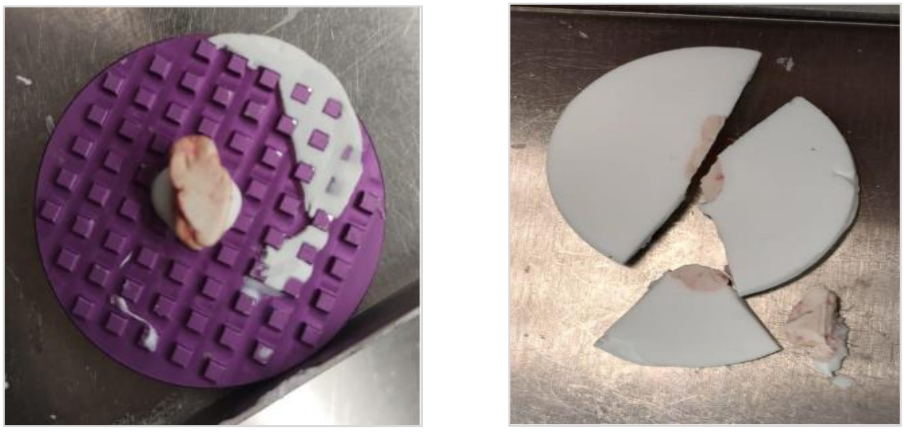


Figure 2.7 Photo of frozen tissue after freezing using PrestoCHILL (A) *Cryosectioning with minimal embedding medium. Tissue sample adhered to the chuck with a small amount of MCC added after the tissue was frozen using PrestoCHILL.* ***(B)*** *Cryosection of sample embedded in MCC. This caused cracking of the tissue and surrounding medium.*

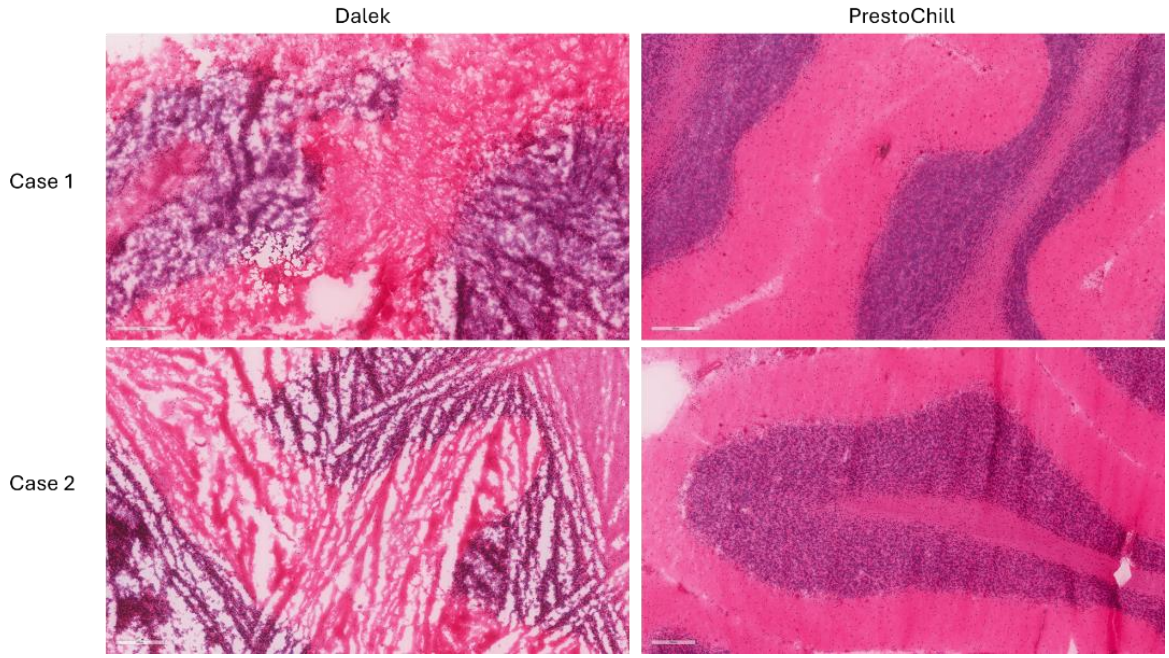


Figure 2.8 H&E images from tissue frozen using Liquid nitrogen vapour compared to PrestoCHILL. Top/bottom left: H&E stained cryo sections of cases frozen in liquid nitrogen vapour (Dalek), showing ice crystal artifact on both sections. **Top/bottom right:** H&E stained cryo sections of the same cases frozen using PrestoCHILL, there was no evidence of ice crystal artefact. * Dalek refers to tissue frozen in a cryoshipper which is pre-chilled with liquid nitrogen vapour.

2.4.3 Comparison of Monarch® HMW DNA Extraction Kit for Tissue and Zymo Quick-DNA Microprep Kit

The DNA purity was assessed using 260/280 ratio. The 260/280 absorbance ratio for DNA extracted from tissue using Monarch kit varied between 1.83 and 1.87 and DNA extracted from tissue using Zymo kit varied between 1.8 and 2.0. The values were found to be normally distributed (D'Agostino Pearson omnibus normality test, p-value Zymo kit 0.4455, Monarch kit 0.5756) and there was no statistically significant differences of 260/280 ratio between DNA extracted from both kits (paired t-test, p value 0.0875) (Figure 2.9).

DNA concentration was measured using qubit fluorometer. The concentration for DNA extracted using the Zymo kit ranged between 23ng/µl and 108ng/µl and for Monarch kit ranged 33ng/µl and 112 ng/µl. Following a D'Agostino Pearson omnibus normality test, the concentration was found to be normally distributed (p value Zymo 0.1722, Monarch 0.3751). There was no significant difference between DNA concentration from both kits (paired t-test, p value 0.7891) (Figure 2.9).

TapeStation was used to assess the fragment size of four out eight of case. The fragment size measure in base pairs (bp) of DNA extracted using the Zymo kit ranged between 18,079 and 25,123b and for Monarch kit ranged between 26,981 and 36,892. Following a Shapiro-Wilk normality test, the fragment size was found to be normally distributed (p value Zymo 0.0885, Monarch 0.9989). There was a significant difference between DNA fragment size from both kits (paired t-test, p value 0.0015) (Figure 2.10).

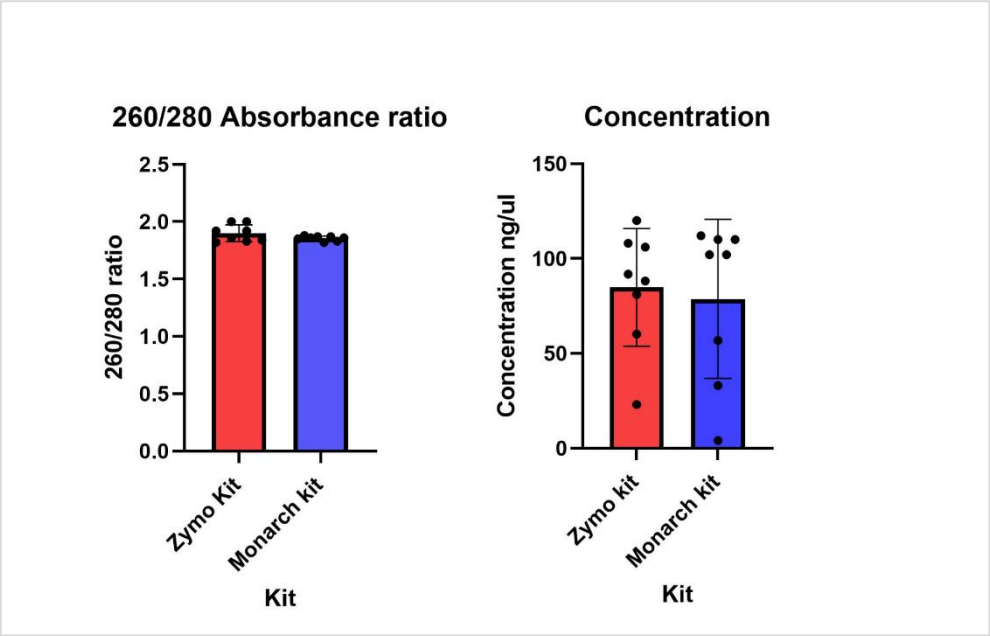


Figure 2.9 Comparison of nucleic acid purity and concentration obtained using Zymo and Monarch extraction kits. Left: *The 260/280 absorbance ratio was measured to assess nucleic acid purity with values between 1.8-2.0 indicating high purity.* ***Right:*** *The concentration of the extracted nucleic acids is show in ng/µl. Bars represent mean values and bars indicate standard deviation.*

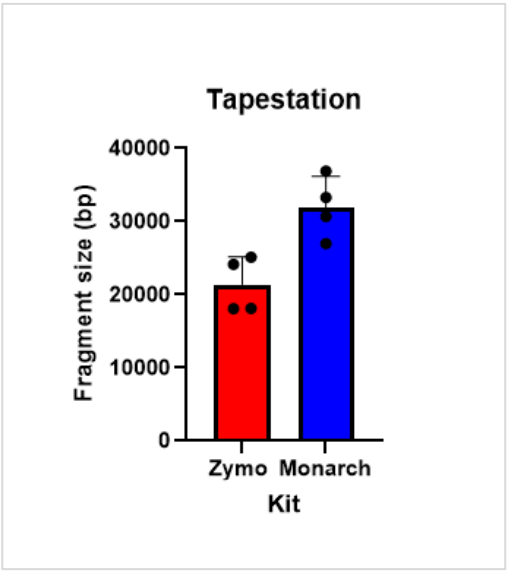


Figure 2.10 Comparison fragment size using Zymo and Monarch extraction kits. *Bars represent mean values and bars indicate standard deviation. Fragment size is measure in base pairs (bp).*

2.5 Discussion

The histological assessment initially indicated that tissue preserved in Hibernate-A supplemented with B27 supplement was of better quality than tissue preserved in TissueReady. However, it was thought that the reason the tissue in TissueReady degraded more rapidly was due to the time delay between fresh dissection and samples being placed into TissueReady gel which could have caused drying out of the tissue (there was no delay for Hibernate-A). Once this possibility was identified, samples were placed directly into TissueReady and these samples showed less degradation compared to previous samples, meaning that there was very little difference in histological tissue quality when comparing the two preservation media.

Statistically, there was only one significant difference between the 260/280 absorbance ratios for DNA extracted from control tissue and stored in Hibernate-A for 96hrs at RT. The 260/280 absorbance ratio of the control tissue was 1.75 and 1.7 which is slightly lower than what can be described as pure DNA. Whereas the absorbance ratios when tissue had been stored in Hibernate-A for 96hrs at RT were 1.84 and 1.8 which is considered pure DNA without RNA or protein contamination (Gallagher, 1998). There were no statistically significant difference between DNA concentration, however, DNA concentration decreased the longer samples were stored in Hibernate-A at both temperatures.

Hibernate-A is a cell culture medium designed for the maintenance of adult neural tissue which can be stored for up to one month at 4°c when supplemented with B-27 supplement and is suitable as a transport media to ship tissue and biological specimens (Brewer and Price, 1996). Following the original publication of Hibernate-A suitability as a storage medium, there was very little in the literature to show the feasibility of Hibernate-A as a

transport medium. More recently a protocol for rapid post-mortem cell culture of diffuse Intrinsic pontine glioma (DIPG) (Lin and Monje, 2017) recommended Hibernate-A as the transport medium for best sample viability. Samples were placed directly into cooled Hibernate-A (supplemented with antibiotics) and shipped overnight on dry ice. This protocol allowed for viable cell culture following overnight transport. A second paper was published in 2018 (Olah *et al,* 2018) where post-mortem brain samples were again stored in cooled Hibernate-A and shipped overnight at 4°c. These samples were suitable for transcriptomic analysis. However, there were no other papers that assessed the quality of tissue following storage in Hibernate-A for over 24hrs at room temperature or 4°c.

Atelerix tissue-ready encapsulates tissues in a gel medium which is able to stabilise the cell membrane and allows tissue samples to be stored at room temperature for up to two weeks (Atelerix, 2025). Atelerix currently have several commercially available products including BeadReady for suspended cells, WellReady for plated cells and TissueReady for primary tissues. BeadReady has been shown to successfully preserve cells cultured from Glioblastoma tissue (Atelerix, 2021) and TissueReady to preserve murine tissues. However, it has not been previously shown if this technology is suitable for preserving human brain biopsy or post-mortem material.

The results suggest that both Hibernate-A and TissueReady are suitable for storage of tissue for up to 96hrs at RT. But it is difficult to draw a conclusion of which novel culture medium resulted in superior preservation of tissues and molecules compared to existing methods such as freezing or fixation due to the small sample size (n=8).

TissueReady requires the use of a tissue culture medium to dilute the hydrogel. In this study, we used Hibernate-A as the culture medium, so combining the two media. However,

Hibernate A may not be the ideal 'partner' medium for the Atelerix gel (particularly as our data suggest that there is a decrease in tissue preservation with Hibernate-A when used in isolation.

Cryopreservation of surgical tissue is important in a clinical setting as standard methods for preserving tissue such as formalin fixed paraffin embedding may make the tissue unsuitable for molecular studies that incompatible with this preservation type. However, conventional freezing methods such as freezing in precooled isopentane or liquid nitrogen vapour not only exposes the user to harmful chemicals and is a health and safety risk, but these methods also result in ice crystal artefact which consequently damage cell morphology which can compromise subsequent analysis. Therefore, rapid freezing is critical for preserving the tissue architecture. The data in this study demonstrates that PrestoCHILL enables rapid and effective freezing of tissue samples with all samples frozen within 60 seconds. This freezing time is faster than conventional methods using liquid nitrogen vapour or pre-cooled isopentane.

Histological assessment of cryosections revealed that ice crystal artefacts were absent in 100% of samples frozen using PrestoCHILL, and cell morphology was well preserved (Figure 2.8). In comparison with tissue frozen in liquid nitrogen vapour which exhibited ice crystal formation (figure 2.8) which can be caused by the slow transition of water from a liquid to a solid state Mckenzie *et al,* 2024. The tissue architecture remained intact following storage for one week at -80°c, indication that PrestoCHILL samples maintain their histological integrity during short-term storage.

The effect of tissue embedding medium on freezing quality was also assessed. Samples that were imbedded partially or without MCC or OCT were easier to handle and remained stable

during cryosectioning, whereas fully embedded tissue displayed poor adherence to the tissue and often suffered post freezing cracking (Figure 2.7). These observations suggest that minimum embedding media improves the ease of sectioning without compromising the quality of the tissue.

DNA extracted from tissue frozen with PrestoCHILL exhibited 260/280 purity ratios between 1.83 and 1.84 which is considered pure DNA (Gallagher, 2001) and yields ranging between 3.3 µg and 7.4 µg. These yields are sufficient for input into Oxford Nanopore ligation sequencing kit, demonstrating that rapid freezing does not adversely affect the nucleic acid quality.

Overall, the results highlight practical advantages of PrestoCHILL over conventional freezing methods. The rapid freezing process preservers tissue morphology and improves the freezing workflow by limiting exposure to harmful chemicals. However, this study only evaluated the short-term storage of the tissue and the stability of the tissue beyond 1 week is untested. Future research should evaluate the performance of this freezing technique to assess long-term storage effects.

Finally in this chapter, this study compared DNA quality and yield from tissue samples extracted using the Monarch and Zymo DNA extraction kits, focusing on purity, concentration, and fragment size. Overall, both kits produced DNA of similar purity and concentration, however a significant difference was observed in DNA fragment size, which suggests kit-specific effects of DNA fragmentation.

The 260/280 absorbance ration from both extraction methods fell within the range that is accepted as pure DNA (1.8-2.0 (Gallagher, 2001). Abnormal 260/280 ratio usually indicate contamination by reagents used in the extraction protocols. As the absorbance ratios fall

into the acceptable range which indicates minimal contamination and there is no significant difference between both kits, this suggests that both extraction protocols are equally effective at removing contaminants.

Similarly, DNA concentration measurements obtained using the Qubit fluorometer showed not significant difference between the two kits. Both kits yielded DNA concentrations within the ranges suitable for downstream molecular applications such as ONT sequencing. These two findings indicate that neither kit offers a clear advantage in terms of DNA yield and purity.

However, analysis of DNA fragment size using tapestation revealed a statistically significant difference between the two extraction methods. DNA extracted from the Monarch kit had a larger fragment size compared to DNA extracted from the Zymo kit suggesting better preservation of high molecular weight DNA. Monarch HMW DNA extraction kit is designed to extract high molecular weight (HMW) genomic DNA by preserving long DNA fragments that are precipitated onto glass beads and has been validated by ONT for use in its ultra-long DNA sequencing library preparation protocol (Oxford Nanopore Technologies, Ultra-long DNA sequencing kit V14, 2025). In comparison, the Zymo kit is not specifically designed for isolation of high molecular weight DNA. This kit relies on silica membrane spin columns, which require repeated high-speed centrifugation steps which can cause mechanical shearing of DNA fragments (Jaudou *et al,* 2022). It is important to note that DNA fragment size analysis was only carried out on a subset of samples (n=4).

However, for downstream application carried out in the chapters below, both kits yield fragment sizes suitable for ONT whole genome sequencing, but the Monarch kit may be favoured when high-molecular weight DNA is required.

2.5.1 Main Conclusions

- **Tissue-preservation:** Both Hibernate-A and TissueReady maintain tissue and DNA integrity for up to 96 hours at room temperature when tissue is handled promptly, with little difference in histological quality.
- **Rapid cryopreservation:** PrestoCHILL enables fast freezing, preserves morphology without ice crystal artifact and produces DNA suitable for downstream applications providing a safer and more efficient alternative to conventional freezing methods.
- **DNA extraction techniques:** Both Monarch and Zymo kits yield DNA of comparable purity and concentrations. However, the Monarch kit better preserves high molecular weight DNA making it the preferable kit for application requiring long fragments.

Chapter 3 Towards real time diagnostics and monitoring of brain tumours

3.1 Aims of chapter

Inter-lab validation is important, and we were one of the first labs to adopt the ROBIN protocol. Therefore, one aim of this chapter is to develop infrastructure in Oxford to provide rapid integrated histomolecular diagnostics of gliomas and implement real time diagnostics.

ROBIN protocol has been demonstrated on intraoperative tissue; another aim of this chapter is to assess if the protocol will work on difficult tissues such as postmortem and does the PM material reflect the diagnosis obtain in vivo with different methods (SR WGS and EPIC array).

3.2 Introduction

Rapid nanOpore Brain Intraoperative classification: ROBIN is an ONT long-read analytical pipeline that integrates methylation based classification along with CNV profiling, SV calling, MGMT promoter methylation and mutation detection to provide an ultra-rapid tumour diagnosis (Deacon *et al,* 2025). This pipeline uses three methylation classifiers; Sturgeon (Vermeulen *et al*, 2023), NanoDX random forest (RF) from RapidCNS2 (Euskirchen *et al,* 2017; Kuschel *et al,* 2021; Patel *et al,* 2022), and CrossNN (Yuan *et al,* 2024).

Sturgeon is a patient agnostic learned neural network trained on simulated nanopore runs from the Capper *et al* (2018a) data set of 2801 methylation profiles from CNS tumours and control samples. Neural networks are a machine learning algorithms modelled on the concept of the human brain and nervous system (Han, Kim, Kim and Youn, 2018). Machine learning nodes which are like the neurons in the brain are able to receive input data from other nodes, process the data and provide an output depending on certain thresholds. Sturgeon can accurately diagnose tumours from low coverage ONT methylation profiles within 40 minutes of starting sequencing (Vermeulen *et al*, 2023).

NanoDX random forest (RF) developed by Euskirchen *et al* (2017) and validated in a cross-laboratory pilot study by Kuschel *et al* (2021). This is a random forest machine learning approach in which multiple decision trees are created and the combined outputs from these trees are used to classify data. This classifier was trained using Heidelberg reference cohort of brain tumour methylation profiles with data generated from the Illumina BeadChip 450K array using the (Capper *et al,* 2018a; Kuschel *et al,* 2021).

CrossNN, like sturgeon is a neural network machine learning algorithm that can classify tumours from methylation profiles. Unlike Sturgeon and RF classifier that can classify from

low-coverage nanopore sequencing, this classifier can be used with sequencing data obtained from different sequencing platforms including ONT and Illumina 450K and EPIC arrays (Yuan *et al,* 2024). CrossNN was used to train two classifiers, a brain tumour classifier that is trained on the Heidelberg brain tumour classifier reference data set and the second classifier is a pan-cancer model and was trained on 178 tumour types across multiple organ sites. Within the ROBIN analytical pipeline, the classifiers are referred to as NanoDX and PanNanoDX. NanoDX is the end-to-end diagnostic pipeline which utilises crossNN which is the machine-learning model responsible for the brain and pan-cancer classification.

The full list of methylation classes included in the Sturgeon classifier is published by Vermeulen *et al* (2018) in Supplementary table 2 and 9. For NanoDX the full list is provided by Capper *et al,* (2018) in supplementary table 1 and Kushel *et al*, (2021) in table 1. For PanNanoDX the list is published by Yuan *et al,* (2025) in supplementary table 2. The lists are available in the cited publications and are not included in the supplementary data of this thesis.

Rapid-CNS^2 is a molecular diagnostic workflow that can provide copy-number profiling, methylation analysis and mutation detection using ONT sequencing data (Patel *et al,* 2022). This workflow uses adaptive sequencing which allows for enrichment of regions of interest by selecting which DNA strands are sequenced. When sequencing using adaptive sampling, a BED file containing regions of interested is selected. With live basecalling, the sequence of the first 100-200 bases can be called and if it is determined that the sequence is not a region of interest, the pore can eject the strand freeing up the pore for the next strand. Rapid-CNS^2 uses regions of interest from the brain tumour NGS panel which consists of 130 genes that are altered in brain tumours (Sahm *et al*, 2016).

ROBIN pipeline has been tested on 50 prospective intraoperative cases in Nottingham, and the classifiers showed a concordance rate with the standard of care integrated diagnosis of 90% (Decon *et al,* 2025). However, inter-lab validation is important to ensure reproducibility and reliability of results. Therefore, this chapter will assess if ROBIN analytical pipeline can successfully be implemented in Oxford and correctly classify brain tumour entities on both intraoperative cases but also to assess if the pipeline will work on more difficult tissues such as post-mortem tissue.

3.3 Methodology

For retrospective ONT sequencing, 6 cases with frozen tissue available from human post-mortem brain tissue and 4 cases with extracted DNA from surgical biopsies were selected for sequencing. Cases were selected that had an integrated histomolecular diagnosis as part of stand of care. For prospective cases, 4 cases with fresh tissue from surgical biopsies were selected, the workflow for prospective cases is shown Figure 3.1. Case demographics are described in Table 3.1.

Table 3.1Case selection for cases for real-time diagnostics and monitoring of brain tumours.

Case ID	**Clinical Diagnosis**	**PM or Surgical**	**Starting material (Fresh tissue, frozen tissue, DNA)**	
NP006-2017	Medulloblastoma	PM	Frozen tissue	Retrospective
NP051-2017	Atypical Teratoid Rhabdoid Tumour (ATRT)	PM	Frozen tissue	Retrospective
NP018-2018	Astrocytoma	PM	Frozen tissue	Retrospective
NP026-2019	Atypical Teratoid Rhabdoid Tumour (ATRT)	PM	Frozen tissue	Retrospective
NP039-2020	Glioblastoma (GBM), IDH wild-type, RTK III	PM	Frozen tissue	Retrospective
NP063-2022	Pilocytic Astrocytoma	PM	Frozen tissue	Retrospective
SH1343-2020	Oligodendroglioma, IDH mutant, 1p19q co-deleted	Surgical biopsy	DNA	Retrospective
SH214-2021	Langerhans cell histiocytosis (LCH)	Surgical biopsy	DNA	Retrospective
SH451-2023	Low grade glioma (LGG) MYB-MYBL1, PSD3::CPNE3 fusion	Surgical biopsy	DNA	Retrospective
SH576-2023	Mesenchymal tumour, morphology compatible with Rhabdomyosarcoma	Surgical biopsy	DNA	Retrospective
SH894-2025	Ependymoma	Surgical biopsy	Fresh tissue	Prospective
SH912-2025	Astrocytoma, IDH mutant	Surgical biopsy	Fresh tissue	Prospective

SH946-2025	Low grade glioma	Surgical biopsy	Fresh tissue	Prospective
SH1049-2025	Astrocytoma, IDH mutant	Surgical biopsy	Fresh tissue	Prospective

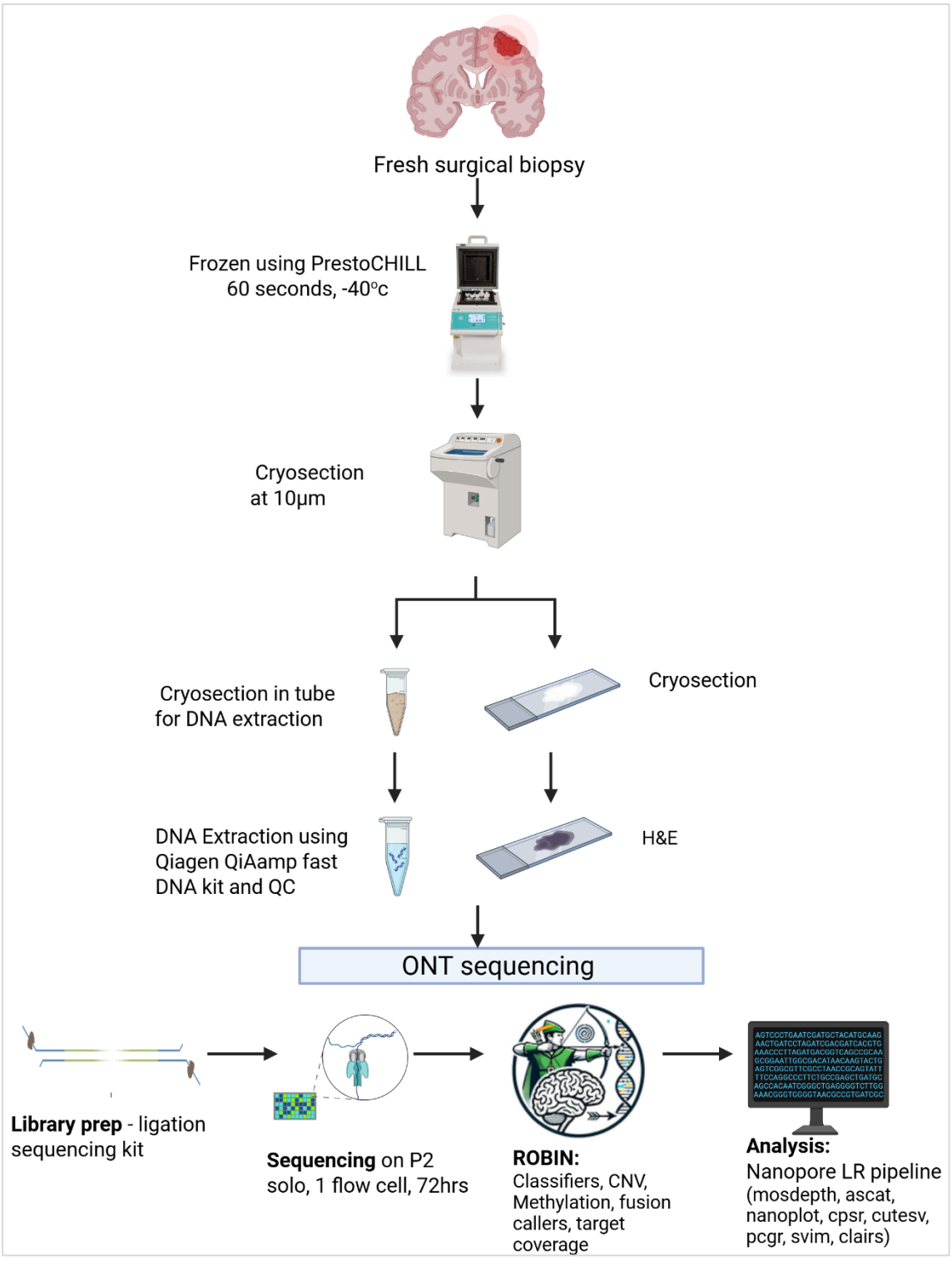


Figure 3.1 prospective ONT sequencing workflow. *A fresh surgical biopsy is received in the lab and frozen using PrestoCHILL. A cryosection is cut from the frozen tissue stained with H&E to assess the quality of the tissue and tumour content. If the tumour content is >25% cryosections are cut for DNA extraction. DNA is extracted using Qiagen QiAamp fast DNA tissue kit, QC is performed using Qubit dsDNA HS Assay kit. 1 µg or 2 µg (depending on yield following DNA extraction) is input into the ONT ligation sequencing kit for DNA library preparation. Libraries are sequencing on a P2 flow cell for 72 hrs. BAM files that are generated from the ONT sequencing are inputted into the ROBIN pipeline in real-time. Following the completion of sequencing (72hrs) sequencing data is analysed using the Nanopore LR pipeline. * SH1049/2025 was not frozen using PrestoCHILL but instead placed in TissueReady when received fresh and stored for 72hrs. DNA was extracted from fresh tissue.*

3.3.1 DNA extraction using Qiagen QiAamp fast DNA kit and quality control

DNA was extracted from 30 mg of frozen post-mortem brain tissue and x40 10 µm cryosections from prospective surgical cases using the QIAamp Fast DNA Tissue Kit (Qiagen, Germany) following the manufactures protocol. DNA from PM brain tissue was eluted in 100 µl EB buffer and DNA from surgical cases was eluted in 50 µl EB buffer. DNA concentration was measured using Qubit™ 4 Fluorometer and Qubit™ dsDNA HS Assay kit (Q32851;ThermoFisher Scientific).

3.3.2 Library preparation using ONT Ultra-long sequencing kit and sequencing

Library preparation of DNA extracted from PM tissue was performed with 1 µg input of genomic DNA using the Ultra-Long DNA Sequencing Kit (SQK-ULK114, Oxford Nanopore Technologies, UK) using ROBIN, version 2 (Deacon, Cahyani, and Loose, 2024), DNA was tagmented using diluted fragmentation mix provided as part of the ONT kit and purified using AmpureXP DNA binding beads (Beckman Coulter, USA). Library concentration was measured using Qubit™ 4 Fluorometer and Qubit™ dsDNA HS Assay kit (ThermoFisher Scientific, USA) prior to flow cell loading. The library was loaded onto R10.4.1 PromethION flowcell (Oxford Nanopore Technologies, UK). Basecalling was performed whilst sequencing using Dorado (version 7.6.7) which is integrated within MinKNOW (version 24.11.8). Reads were called using High-accuracy (HAC) model (v4.3.0, 400bps) with 5hmC and 5mC modifications. Reads were mapped to GRCh38 during basecalling and the resulting BAM files used for subsequent analysis.

3.3.3 Library preparation using ONT Ligation sequencing kit and sequencing

DNA samples from surgical biopsies were prepared with the ONT Ligation sequencing kit V14 with an adjusted protocol. 2 µg of DNA was repaired and prepared for adapter ligation purified using AmpureXP DNA binding beads (Beckman Coulter, USA). Library concentration

was measured using Qubit™ 4 Fluorometer and Qubit™ dsDNA HS Assay kit (ThermoFisher Scientific, USA) prior to flow cell loading. The library was loaded onto R10.4.1 PromethION flowcell (Oxford Nanopore Technologies, UK). Basecalling was performed whilst sequencing using Dorado (version 7.6.7) which is integrated within MinKNOW (version 24.11.8). Reads were called using High-accuracy (HAC) model (v4.3.0, 400bps) with 5hmC and 5mC modifications. Reads were mapped to GRCh38 during basecalling and the resulting BAM files used for subsequent analysis.

3.3.4 Real time visualization (ROBIN)

For real time visualization of the sequencing data, ROBIN package (0.1.0) (Deacon *et al*, 2025) was used. To run ROBIN, Linux operating system is required, for this Ubuntu 22.04.5 LTS for windows subsystem for Linux was used. To see files were transferred from Windows to Ubuntu, see appendix 1. Confidence thresholds for each classifier were set in the original publications and applied to ROBIN (Sturgeon ≥0.8; CrossNN ≥ 0.2; Rapid-CNS2 RF ≥70). The ROBIN pipeline was executed until all BAM files generated during the 72-hour ONT sequencing run were fully processed. ROBIN configuration file can be found in appendix 2.

```
conda activate robin
robin -c robin_configuration_file.txt \
output_directory_robin_must_be_empty_before_starting -w \
/root/storing_data_for_robin/
```

Table 3.2 Confidence levels of classifiers

Classifier	High confidence	Medium confidence	Low confidence
Sturgeon	>85%	>65%	<65%
NanoDX	>50%	>25%	<25%
PanNanoDX	>50%	>25%	<25%
RF	>85%	>65%	<65%

3.3.5 Long-read Whole Genome Sequencing Analysis

Bioinformatic analysis was performed by Dr Ebony Cave using the following pipeline. Long-read ONT Pipeline, which was developed by Andrew Beggs, University of Birmingham (2024) was used for the whole genome sequencing analysis. A basic overview of the pipeline in shown in Figure 3.2 (See appendix 4 for full code script). Basecalled reads were aligned to Human GRCh38 using minimap2 version 2.27 -r1193 (Li, 2018) using the preset for long-read, high quality for fast and accurate mapping of nanopore data. Samtools was used to merge all BAM files for analysis. Genome wide coverage was assessed using mosdepth version 0.2.6 (Pedersen and Quinlan, 2018) fast mode to calculate average read depth per 1000 base pairs. Structural variants (SV) were called using Savana gene fusion caller version 1.3.1 (Cortes-Ciriano *et al,* 2024) tumour only mode; Sniffles2 version 2.3.3 (Smolka *et al,* 2024); cuteSV version 2.0.2 (Jiang *et al,*2020); SVIM 1.4.2 (Heller and Vingron, 2019); Clair3 version 1.0.7 (Zheng *et al,* 2022). Epigenetic alterations including methylation was detected using modkit

0.4.1 (Oxford Nanopore Technologies, https://github.com/nanoporetech/modkit). Only sites with a fraction of modified reads of > 0.8 were retained for downstream analysis. Clinical reports were generated using CSPR and PCGR reporting version 2.1.2.

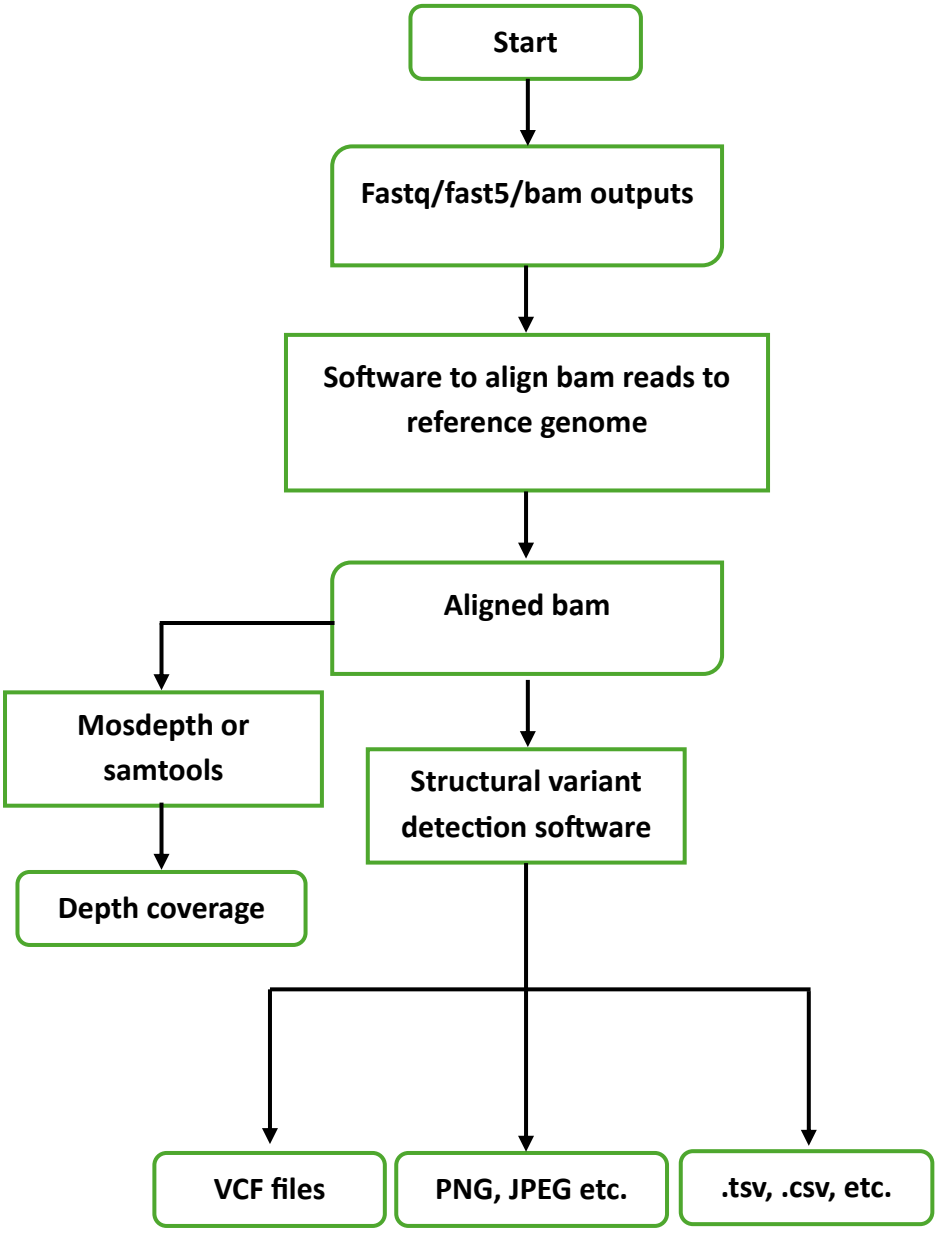


Figure 3.2 Long-read WGS pipeline. *During ONT sequencing, the output of the raw data is in the form of POD5 files. These files containing the electrical signal output are converted into bases using a basecaller software, dorado. This produces FASTQ and BAM outputs which are then aligned to a reference genome. The aligned bam files can be inputted into any analysis software to assess depth coverage and detect variants. The output from these analysis software's is in the format of VCF files or reports.*

3.4 Results

3.4.1 Standard of care testing case review

SH214-2021 Langerhans cell histiocytosis

This case is that of a child with a lytic lesion of the skull bone with the histology of Langerhans cell histiocytosis (LCH). Immunohistochemistry showed the characteristic nodules of CD1A positive tumour and a mixed inflammatory infiltrate (Figure 3.3). At receipt of the biopsy, tissue was fixed and processed into FPPE blocks and a sample frozen for short-read whole genome sequencing. Unstained FFPE sections were sent for BRAF mutation analysis which detected a mutation at BRAF p.Val600. The preliminary histological report was available 9 days after biopsy sample receipt, and the final integrated diagnosis was reported at 29 days. In addition, a sample was sent for short-read, whole genome sequencing which confirmed the BRAF p.Val600 mutation but no other variants were detected. Although the data received from short-read sequencing was not new clinically relevant data, the time from receipt of biopsy to short-read report being available was 257 days.

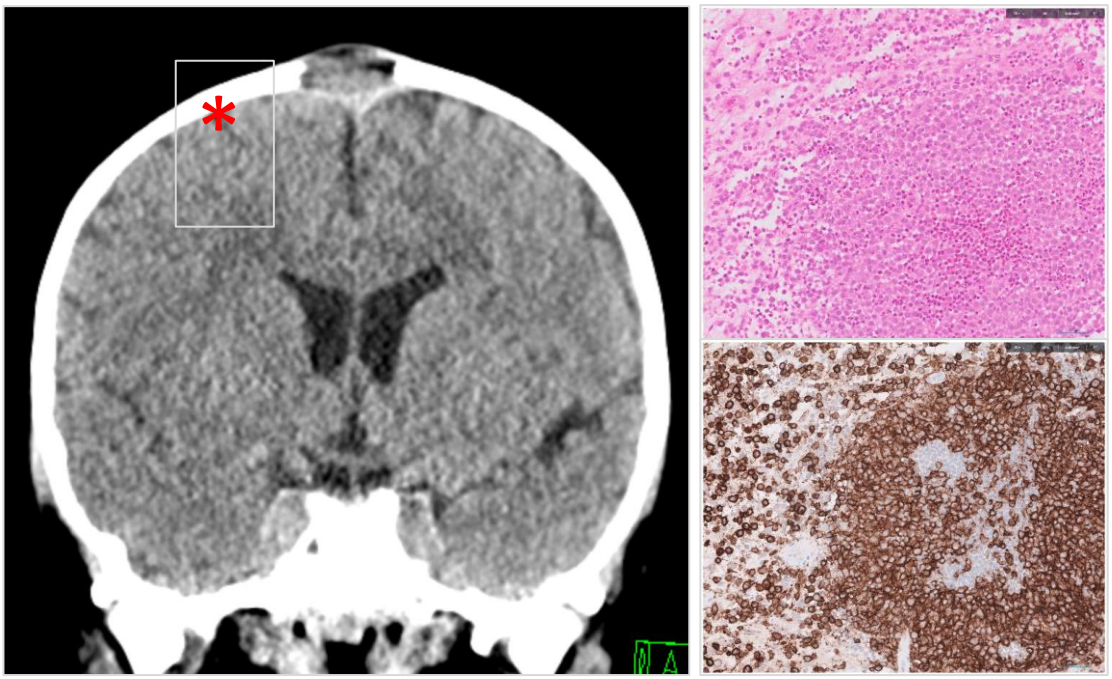


Figure 3.3 SH214-2021, Langerhans cell histiocytosis*. A lytic lesion of the skull bone (above the red asterisk) in a child with the histology of Langerhans cell histiocytosis (LCH) characterised by modules of CD1A-positive tumour cells and a mixed inflammatory infiltrate.*

SH451-2023 Low grade glioma – MYB/MYBL1 altered

This case is that of a left frontal intrinsic lesion with histological analysis demonstrating mildly hypercellular atypical glial cells with a low proliferation rate (MIB1 ≈ 1%) (Figure 3.4). At receipt of the biopsy, tissue was fixed and processed into FPPE blocks and a sample frozen for short-read whole genome sequencing. Figure 3.5 illustrates FFPE tissue use, with 25 sections used for immunohistochemistry, which cannot achieve an integrated diagnosis. The histological diagnosis, which was available within 7 days of receipt of the biopsy, was a low-grade glioma and may represent a paediatric type diffuse low-grade glioma with MAPK pathway altered or MYB- or MYBL1 altered. However, molecular characterisation is required to confirm this diagnosis. This case received methylation profiling using Illumina EPIC array. The tumour was classified with high confidence (0.99871) as low grade glioma, MYB/MYBL1 altered. The final integrated diagnosis was available within 45 days. A sample was sent for short-read whole genome sequencing and was consistent with a previously reported finding

of an inversion in chromosome 8 affecting *MYBL1* gene and in addition, a previously unreported *PSD3::CPNE3* gene fusion was confirmed. The short-read results included new clinically relevant data but was only available 211 days after receipt of the biopsy.

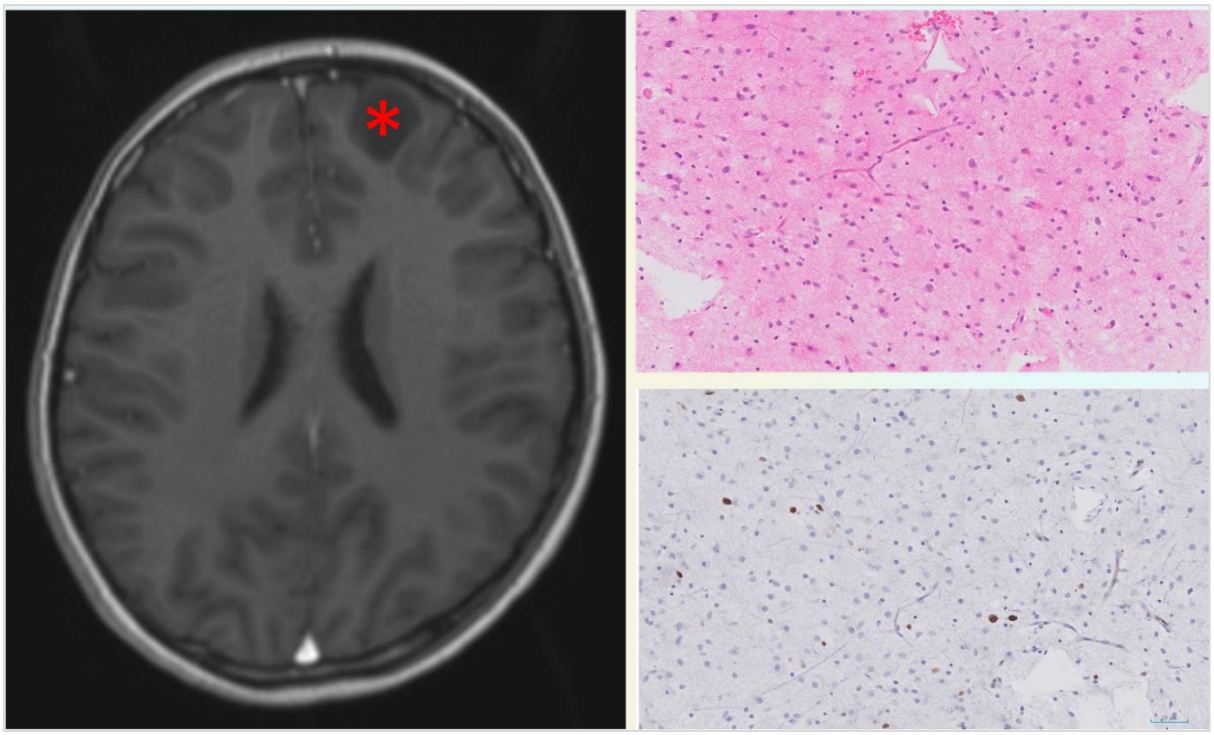


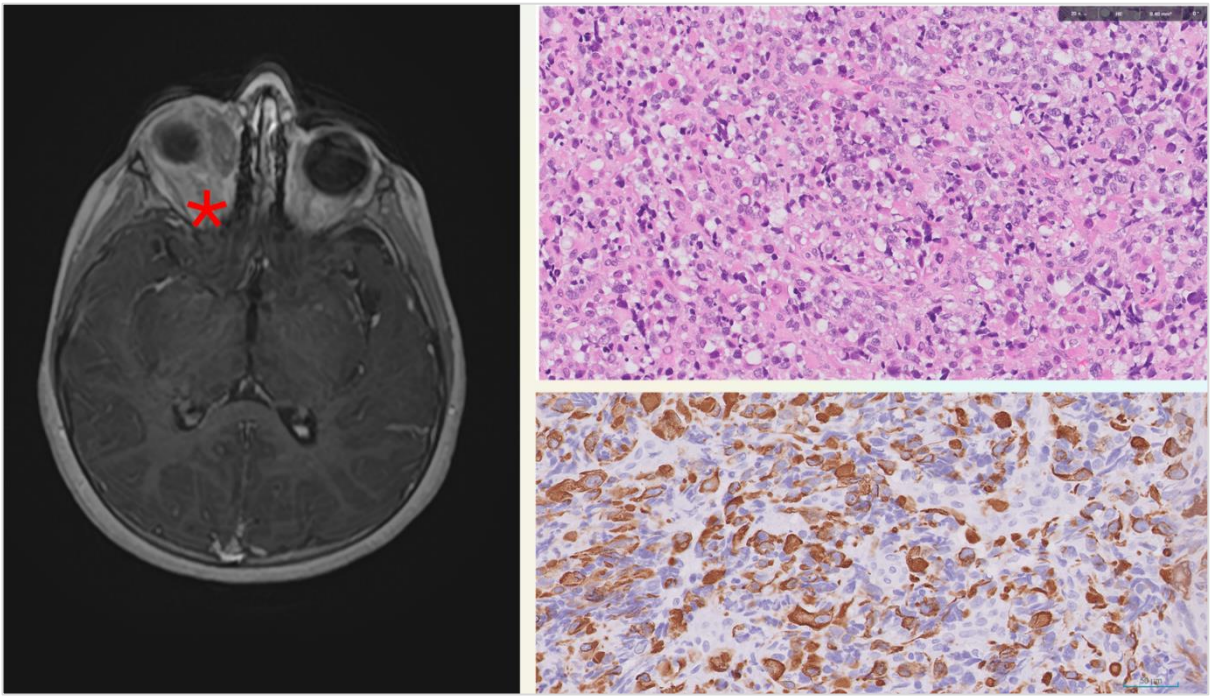
Figure 3.4 SH451-2023, Low grade glioma, MYB/MYBL1 altered. *A hypointense left frontal intrinsic lesion (asterisk) with indistinct, mildly hypercellular atypical glial cells with a low proliferation rate (MIB1). A morphological diagnosis is not possible.*

Figure 3-5 SH451-2023 Illustration of tissue use and iterative immunohistochemistry using standard-of-care approaches that cannot achieve a state-of-the art diagnosis. This tumour type can only be diagnosed using epigenomic and genomic analyses.

SH576-2023 Rhabdomyosarcoma

The MRI from this case revealed an ill-defined retrobulbar abnormality of the right orbit. Histological examination revealed a mesenchymal tumour, expressing markers of myocyte lineage, which are consistent with rhabdomyosarcoma (Figure 3.6). At receipt of the biopsy, tissue was fixed and processed into FPPE blocks and a sample frozen for short-read whole genome sequencing. A preliminary histological report was available after 7 days of receipt of the biopsy. A final integrated diagnosis was reported 43 days after biopsy which included immunohistochemistry and molecular cytogenetic testing to assess the presence of an FOXO1 rearrangement, which was not detected. Finally, a frozen sample was sent for short-read whole genome sequencing, the result was available 202 days after the initial biopsy.

Figure 3.6 SH576-2023 , Rhabdomyosarcoma. *An ill-defined retrobulbar abnormality of the right orbit*



(asterisk). Histology reveals a poorly differentiated malignant neoplasm expressing markers of myocyte lineage differentiation, consistent with rhabdomyosarcoma.

SH946-2025 Low grade glioma DNT

This case is a heterogeneously enhancing mesial temporal lobe lesion with previous haemorrhage. At receipt of the biopsy, tissue was fixed and processed into FPPE blocks and a sample for research was frozen using PrestoCHILL prior to ONT sequencing. H & E staining of a cryosection frozen by PrestoCHILL showed good cryoarchitecture and no obvious ice crystal artefact (Figure 3.7). Immunohistochemistry performed on 25 slides (Figure 3.8), was insufficient to achieve an integrated diagnosis for this tumour type, therefore molecular testing is required. A preliminary histological diagnosis was available within 6 days. This case received methylation profiling using Illumina EPIC array. The tumour was classified as low grade glial/glioneuronal – dysembryoplastic neuroepithelial tumour (DNT). The final integrated diagnosis was available within 25 days of receipt of biopsy.

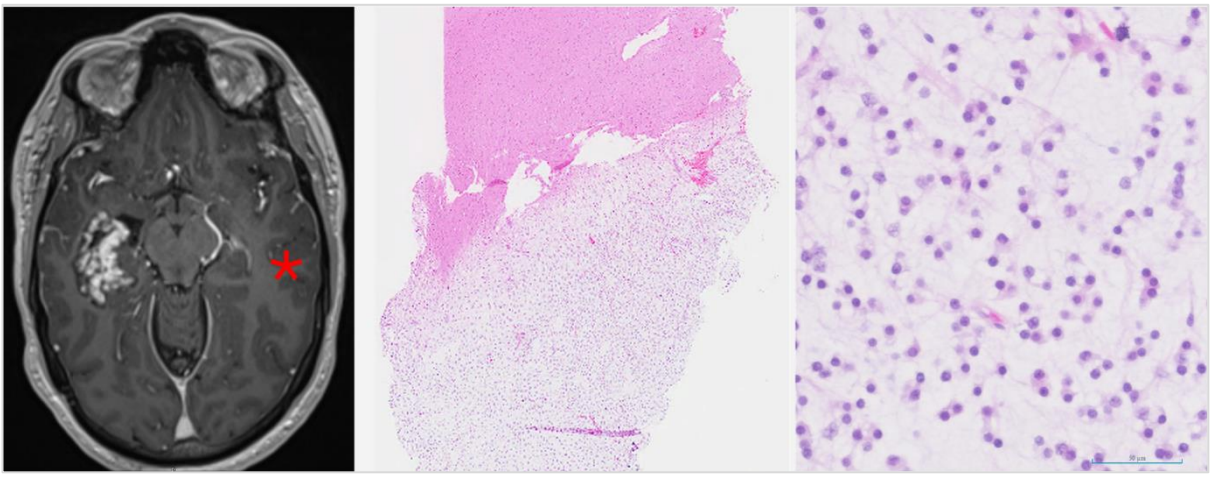


Figure 3.7 SH946-2025 Low grade glioma, DNT. *A heterogeneously enhancing mesial temporal lobe lesion with previous haemorrhage (asterisk). Histology shows a low-cellularity glioneuronal neoplasm with an abrupt border to normal brain (centre) . Histological diagnosis using immunohistochemistry was not possible. Note that a prestoCHILL frozen cryosection cut prior to ONT sequencing shows good cytoarchitecture and no obvious artefact (right).*

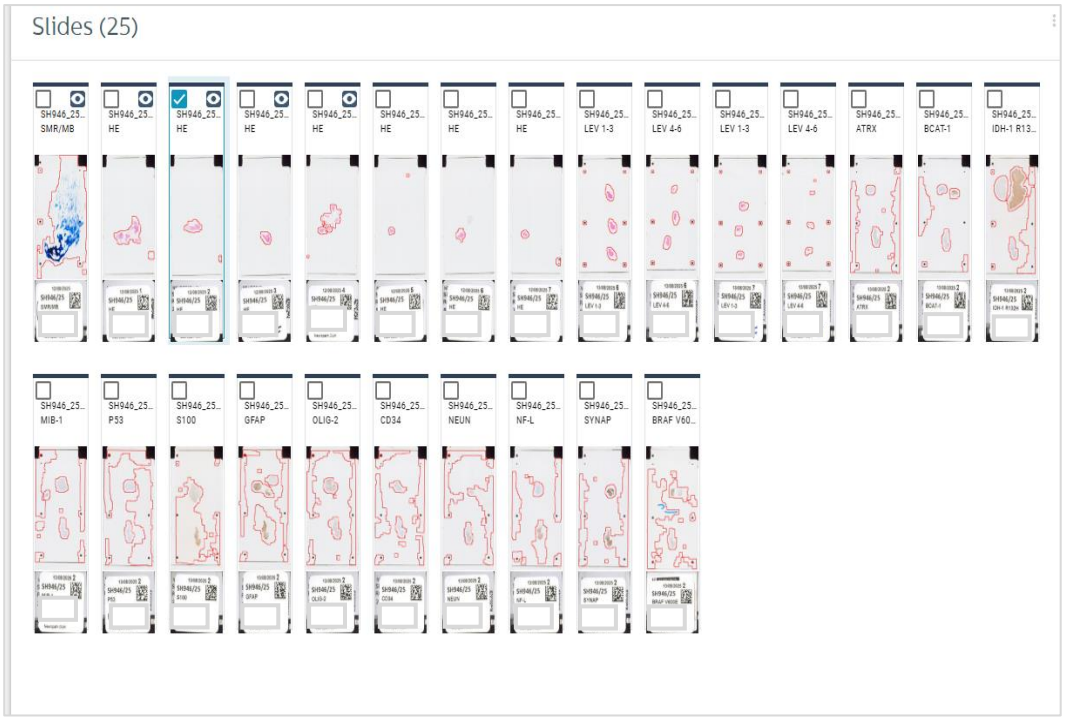


Figure 3.8 SH946-2025 Illustration of tissue use and iterative immunohistochemistry using standard-of-care approaches that cannot achieve a state-of-the art diagnosis. This tumour type can only be diagnosed using epigenomic and genomic analyses.

The sample for research underwent prospective ONT sequencing at the time of surgery. The four methylation classifiers in the ROBIN pipeline (Sturgeon, NanoDX, PanNanoDx, and Random Forest) correctly classified this tumour as a low grade glioma, DNT (Figure 3.9). From receipt of biopsy sample to the correct class being called was 3 hours.

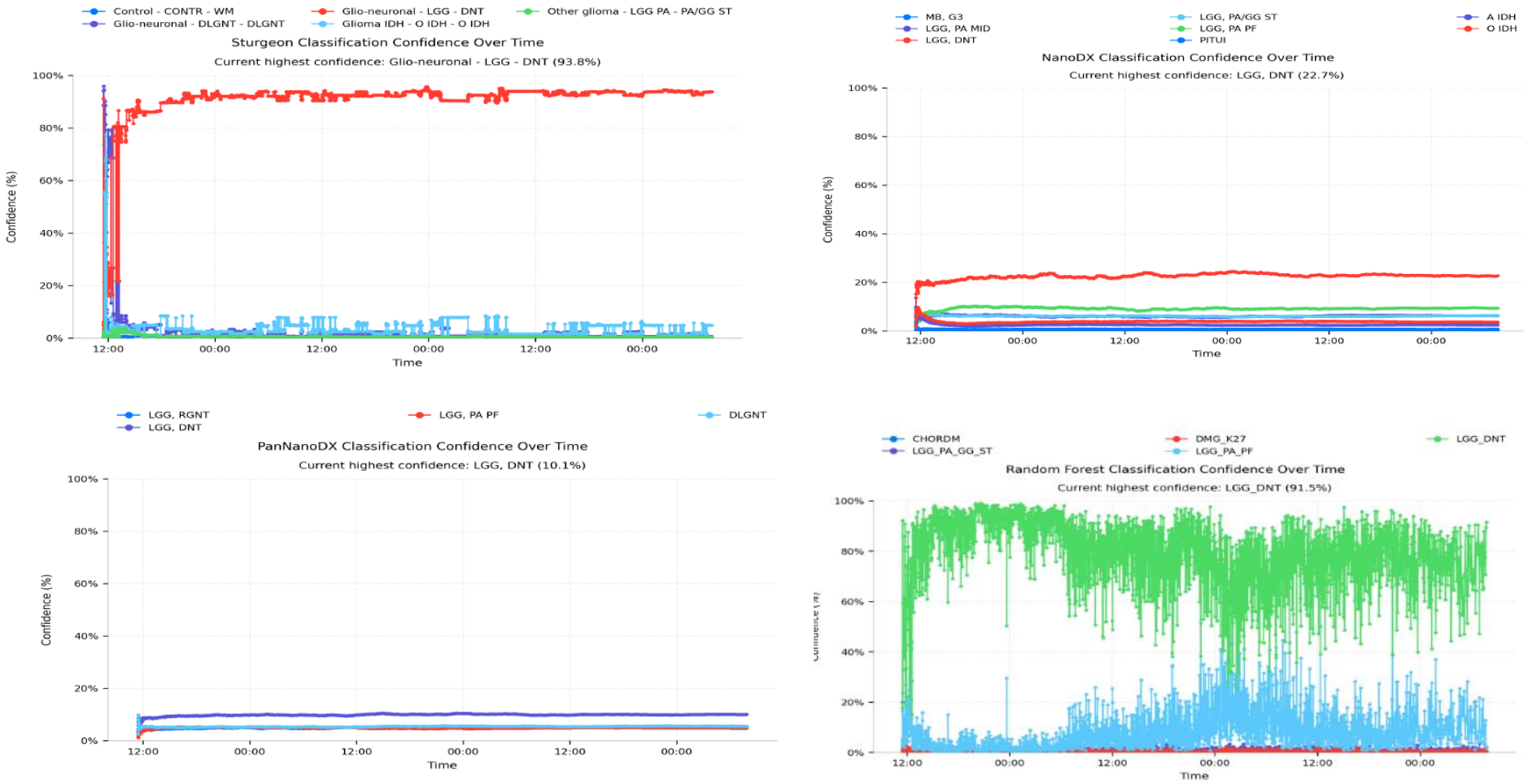


Figure 3.9 Methylation classification results for SH946-2025*. The graphs show classification class and confidence scores over time. The four classifiers correctly classified this case as low grade glioma, DNT.*

3.4.2 ROBIN Methylation classifiers

Of the post-mortem case six cases were retrospectively sequenced, and classification results were validated against standard of care integrated diagnoses (Table 3.3). 100% of cases (n=6) were classified in concordance with standard of care integrated diagnosis within 24 hours of sequencing. The three classifiers (Sturgeon, NanoDx, Pan NanoDX and random forest) predicted the same tumour entity in five out of six correctly classified cases. One case (NP039-2020) was classified by NanoDX and Pan NanoDX as GBM RTK III subtype with confidence levels of 20.9% and 17.6% whereas random forest classifier, classified this as a GBM RTK II subtype with a confidence level of 59.9% (low) and sturgeon classified as GBM – MYCN with a confidence level of 69.2% (medium).

Four surgical biopsy cases with surplus DNA from standard of care short-read sequencing were retrospectively sequenced and classification results were validated against standard of care integrated diagnoses (Table 3.3). 50% of the cases (n=2) were correctly classified by all the classifiers. Case SH214-2021 was incorrectly classified by Sturgeon, NanoDX and random forest as control tissue but correctly classified by Pan NanoDX as LCH - Langerhans cell histiocytosis. SH576-2023 was also incorrectly classified by all three classifiers. The SoC clinical diagnosis was Embryonal Rhabdosarcoma (eRMS), sturgeon classified as Ewing sarcoma family tumour with CIC alteration at high confidence levels, NanoDx classified as GBM RTK III with low confidence levels, and RF classified as meningioma (MNG) with low confidence levels. However, Pan NanoDX classified as SARC – RMS like (Sarcoma – Rhabdosarcoma like) but with low confidence levels. Methylation classification graphs with confidence over time are in Figure. A1 of appendix 6.

Table 3.3 Standard of care integrated diagnosis and ROBIN methylation classification class with confidence scores (PM and retrospective surgical cases).

Case ID	Integrated clinical diagnosis	ROBIN methylation class							
		Sturgeon	Confidence (%)	NanoDX	Confidence (%)	Pan NanoDX	Confidence (%)	RF	Confidence (%)
NP006-2017	Medulloblastoma	Medulloblastoma G3	99.9 (High)	Medulloblastoma G3	91.2 (High)	Medulloblastoma G3	65.6 (High)	Medulloblastoma G3	100 (High)
NP018-2018	Astrocytoma	Glioma – IDH – A	98.6 (High)	A-IDH	39.0 (Medium)	A-IDH	19.7 (Low)	A-IDH	100 (High)
NP026-2019	ATRT – SHH	Embryonal - ATRT – SHH	94.6 (High)	ATRT – SHH	37.2 (Medium)	ATRT – SHH	13.7 (Low)	ATRT – SHH	96.1 (High)
NP051-2017	ATRT	ATRT – MYC	99.8 (High)	ATRT – MYC	70.5 (High)	ATRT – MYC	34.0 (Medium)	ATRT – MYC	98.5 (High)
NP039-2020	GBM RTK III	GBM MYCN	69.2 (Medium)	GBM RTK III	20.9 (Low)	GBM RTK III	17.6% (Low)	GBM RTK II	59.9 (Low)
NP063-2022	Pilocytic Astrocytoma	LGG – PA	99.8 (High)	LGG – PA	41.9 (Medium)	LGG – PA	21.4 (low)	LGG – PA	78.1 (Medium)
SH451-2023	Low grade glioma – MYB	LGG – MYB	98.2 (High)	LGG – MYB	27.2 (medium)	LGG – MYB	12.8 (Low)	LGG – MYB	69.5 (medium)
SH576-2023	Mesenchymal tumour – Rhabdosarcoma	Mesenchymal – EFT – CIC	88.9 (High)	GBM RTK III	5.1 (Low)	SARC (RMS-like)	7.0 (Low)	MNG	4.5 (Low)
SH214-2021	Langerhans cell histiocytosis	Control	94.6 (High)	Control inflammation	23.6 (Low)	LCH	12 (Low)	Control	99.7 (High)
SH1343-2020	Oligodendroglioma, IDH mutant, 1p19q codeleted	Glioma – IDH – O	99.6 (High)	O IDH	62.4 (High)	O IDH	26.0 (medium)	O IDH	100 (High)

Four cases were prospectively sequenced at the time of surgery. Classification results were later validated against standard of care integrated diagnoses.

For this cohort, three of the four sample were received in the lab and assessed by a Neuropathologist prior to freezing using PrestoCHILL for 60 seconds. A cryo section was cut and stained with H&E to assess the quality of the tissue and tumour content.

DNA was extracted from the tissue using Qiagen QiAamp Fast DNA Tissue Kit in 40 minutes including quality control using Qubit Fluorometer. Library preparation using ligation sequencing kit took a total time of 1 hour 40 minutes, loading the flow cell and starting the sequencing run took a further 20 minutes with the total time from sample receipt to sequencing being 2 hour 45 minutes. Within 10 minutes of sequencing, the methylation classifiers were calling tumour class. For all prospective cases, the correct class was called under 5 minutes of running ROBIN (SH894-2025, Ependymal EPN MPE; SH912-2025, Glioma – IDH – A; SH946-2025, LGG DNT; SH1049-2025., Astrocytoma). Figure 3.10 shows the time it took from freezing tissue to start of sequencing. Following sequencing all cases were classified by all methylation based classifiers with varying degrees of confidence in concordance with standard of care diagnosis shown in Table 3.4.

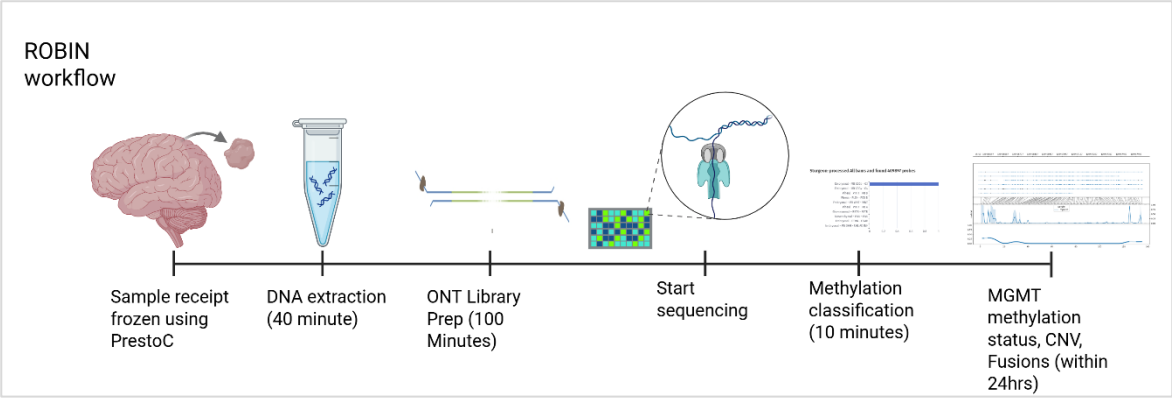


Figure 3.10 Timeline from freezing sample to methylation based classification. *DNA was extracted from frozen tissue within 40 minutes of freezing the tissue sample using PrestoCHILL. ONT library preparation was completed within 1 hour 40 and sequencing commenced with 20 minutes following flow cell flushing and loading. Methylation classifiers were calling classes within 10 minutes of sequencing.*

Table 3.4 Standard of care integrated diagnosis and ROBIN methylation classification class with confidence scores (Prospective surgical cases).

Case ID	Integrated clinical diagnosis	ROBIN methylation classifiers							
		Sturgeon	Confidence (%)	NanoDX	Confidence (%)	Pan NanoDX	Confidence (%)	RF	Confidence (%)
SH894-2025	Ependymoma	Ependymal EPN MPE	98.1 (High)	EPN MPE	62.8 (High)	EPN MPE	34.0 (Medium)	EPN MPE	99.8 (High)
SH912-2025	Astrocytoma	Glioma – IDH – A	94.3 (High)	A IDH	47.5 (Medium)	A IDH	12.4 (Low)	A IDH	95.3 (High)
SH946-2025	Low grade glioma	LGG DNT	93.8 (High)	LGG DNT	22.7 (Low)	LGG DNT	10.1 (Low)	LGG DNT	91.5 (High)
SH1049-2025	Astrocytoma	Glioma – IDH – A	98.3 (High)	A IDH	42.0 (Medium)	A IDH	16.6 (Low)	A IDH	94.6 (High)

3.4.2 MGMT promotor methylation

Of the 14 case sequenced, only six cases received MGMT methylation testing as SoC. Of the six cases that LR analysis could be compared with SoC, five cases (85%) LR results concurred with the SoC methylation analysis results (n=3 unmethylated, n=2 methylated). For one case (NP063-2022) the EPIC chip failed to classify the case and therefore the MGMT promoter methylation was not assessed. This feature of the ROBIN pipeline takes longer than the classifiers to produce a result as the coverage required is higher, however methylation status of the MGMT promoter is still generated within 24 hours of sequencing time. An example of the output from ROBIN pipeline is in Figure 3.11. MGMT methylation outputs are in Figure A.2 of appendix 7.

Table 3.5 Comparison of SoC and ROBIN MGMT promotor methylation status. *SoC methylation status is determined by EPIC array and is compared to methylation status determine by Oxford Nanopore long-read sequencing and ROBIN pipeline. Samples are classified as methylated or unmethylated based on methylation thresholds. Clinically relevant methylation detected by ROBIN pipeline is defined as >25% across the region where as SoC EPIC array is defines methylated as a score above 0.3582.*

CASE ID	SOC METHYLATION STATUS (EPIC)	LR MGMT PROMOTER METHYLATION STATUS	LR MGMT PROMOTER AVERAGE METHYLATION (%)
NP006-2017	Not submitted for EPIC array	Unmethylated	10.6
NP018-2018	Not submitted for EPIC array	Methylated	52.5
NP026-2019	Unmethylated	Unmethylated	8.5
NP051-2017	Not submitted for EPIC array	Unmethylated	1.2
NP039-2020	Methylated	Methylated	69
NP063-2022	EPIC array failed	Unmethylated	6.9
SH451-2023	Unmethylated	Unmethylated	8.6
SH576-2023	Not submitted for EPIC array	Unmethylated	4.8
SH214-2021	Not submitted for EPIC array	Unmethylated	5.8
SH1343-2020	Not submitted for EPIC array	Unmethylated	5.8
SH894-2025	Not submitted for EPIC array	Methylated	53.1
SH912-2025	Not submitted for EPIC array	Methylated	46.2
SH946-2025	Unmethylated	Unmethylated	8.4
SH1049-2025	Methylated	Methylated	51.4

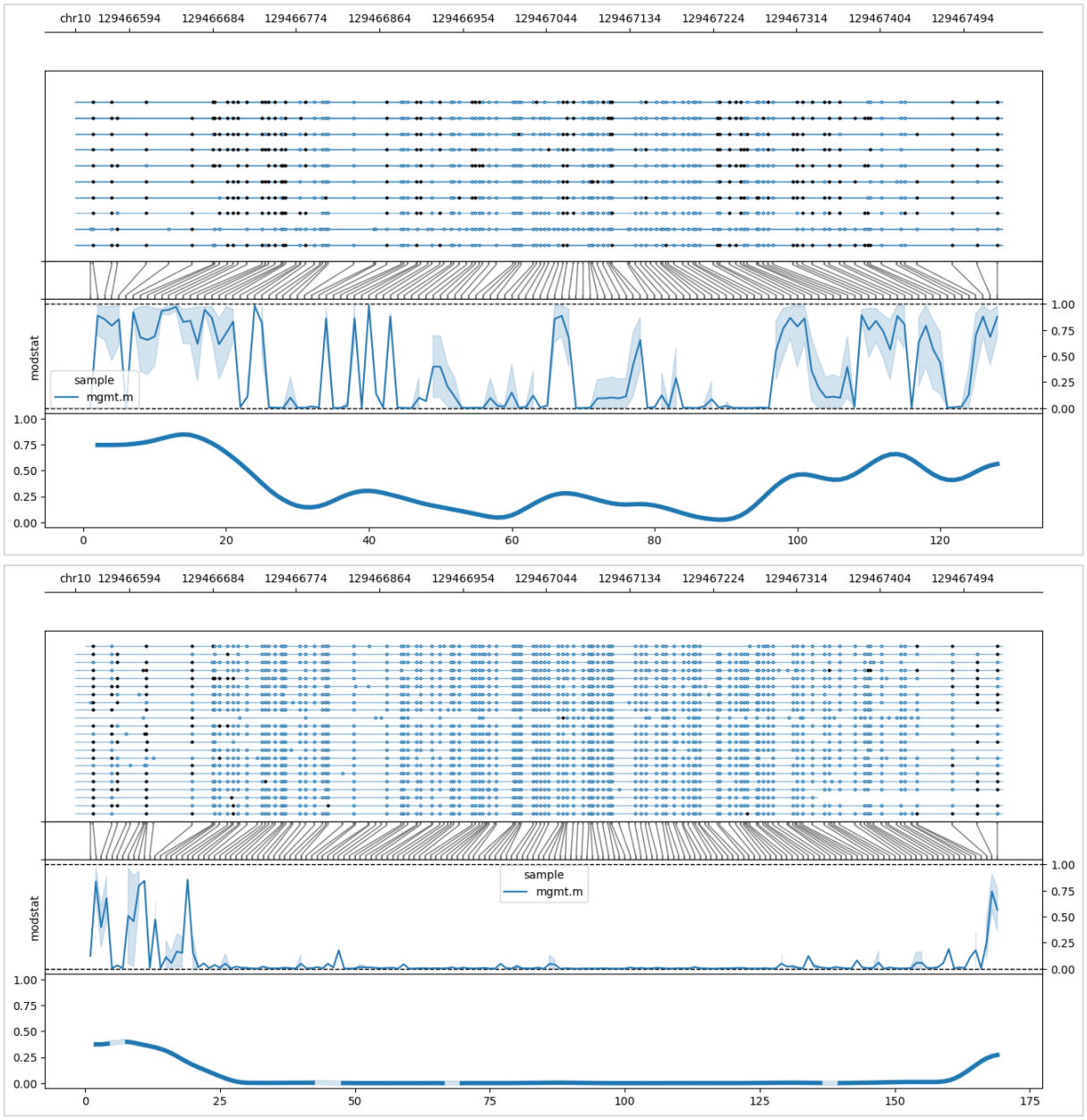


Figure 3.11 Example output of MGMT methylation status generated by ROBIN pipeline. A) Output from an Ependymoma case that is methylated. B) Output from a low grade glioma case that is unmethylated. *MGMT promoter is located on Chromosome 10, and each dot represents a CpG site within that region. The blue lines correspond to a single read. The black dots indicate CpG methylated sites and the blue dots are unmethylated sites.*

3.4.3 Copy number variants, amplifications, deletions and fusions

Copy number profiles, amplifications, deletions and gene fusions generated by ROBIN were in concordance with standard of care testing results. Copy number variants can successfully be detected by ROBIN for example, sample SH1343-2020 is confirmed as an oligodendroglioma and the copy number profile generated showed loss of short chromosome 1 (1p) and long arm of chromosome 19 (19q) as shown in Figure 3.12. Stand of care testing for this case also identified loss of short arm of chromosome 9 (9p) which was displayed on the ROBIN CNV plot.

A key characteristic of atypical teratoid rhabdoid tumour (ATRT) is biallelic inactivation of *SMARCB1*. Two ATRT cases (NP051-2017 and NP026 2019) sequenced and analysed through the ROBIN pipeline and in both case the SMARCB1 deletion was successfully detected. ROBIN outputs boxplots showing target coverage. For both of these cases the box plots show consistent coverage across the chromosomes, while *SMARCB1* on chromosome 22 showed reduce coverage which fell below the whiskers of the boxplots of the chromosome 22 coverage distribution and was identified as a global outlier. To further examine the low cover, read alignment across the gene was visualised in a genome browser which showed sparse coverage across the gene loci, indicating a low number of reads and deletion of *SMARCB1.*

An example of the boxplots showing target coverage and global outliers and Genome browser output is shown below in Figure 3.13.

Figure 3.12 Genome-wide and chromosome level copy number variation (CNV) plots for SH1343-2020. Top: *Genome-wide CNV profile showing copy number differences across all chromosomes. Alternating colours distinguish individual chromosomes.* ***Middle/bottom:*** *Chromosome-specific CNV profiles for chromosome 1, chromosome 19 and chromosome 9.*

Figure 3.13 Target coverage outliers highlighting SMARCB1 on chromosome 22. Top: *Box plot showing target sequencing coverage distribution across all chromosomes. Each box shows the range with median coverage per chromosomes and whiskers shows the range. Red points mark global outliers. SMARCB1 on chromosome 22 shows low-coverage outlier.* ***Bottom:*** *Genome browser (JBrowse 2) view of chr22:23,776,738-chr22:23,847,486 region encompassing SMARCB1. The sparse read coverage across the gene is consistent with the low coverage outlier detected in the box plot, suggesting deletion of this gene.*

Amplification of *MYC* is a common feature of group 3 medulloblastoma. NP006-2017 was classified by ROBIN as Medulloblastoma, G3 and target coverage plots similar to the plots above showed consistent coverage across all chromosomes while *MYC* on chromosome 8 was identified as a global outlier showing increased coverage rising above the whiskers of the boxplots. Again, to further examine the high cover, read alignment across the gene was visualised in a genome browser which showed high coverage across the gene loci, indicting an amplification of *MYC*.

An example of the boxplots showing target coverage and global outliers and Genome browser output is shown below in Figure 3.14.

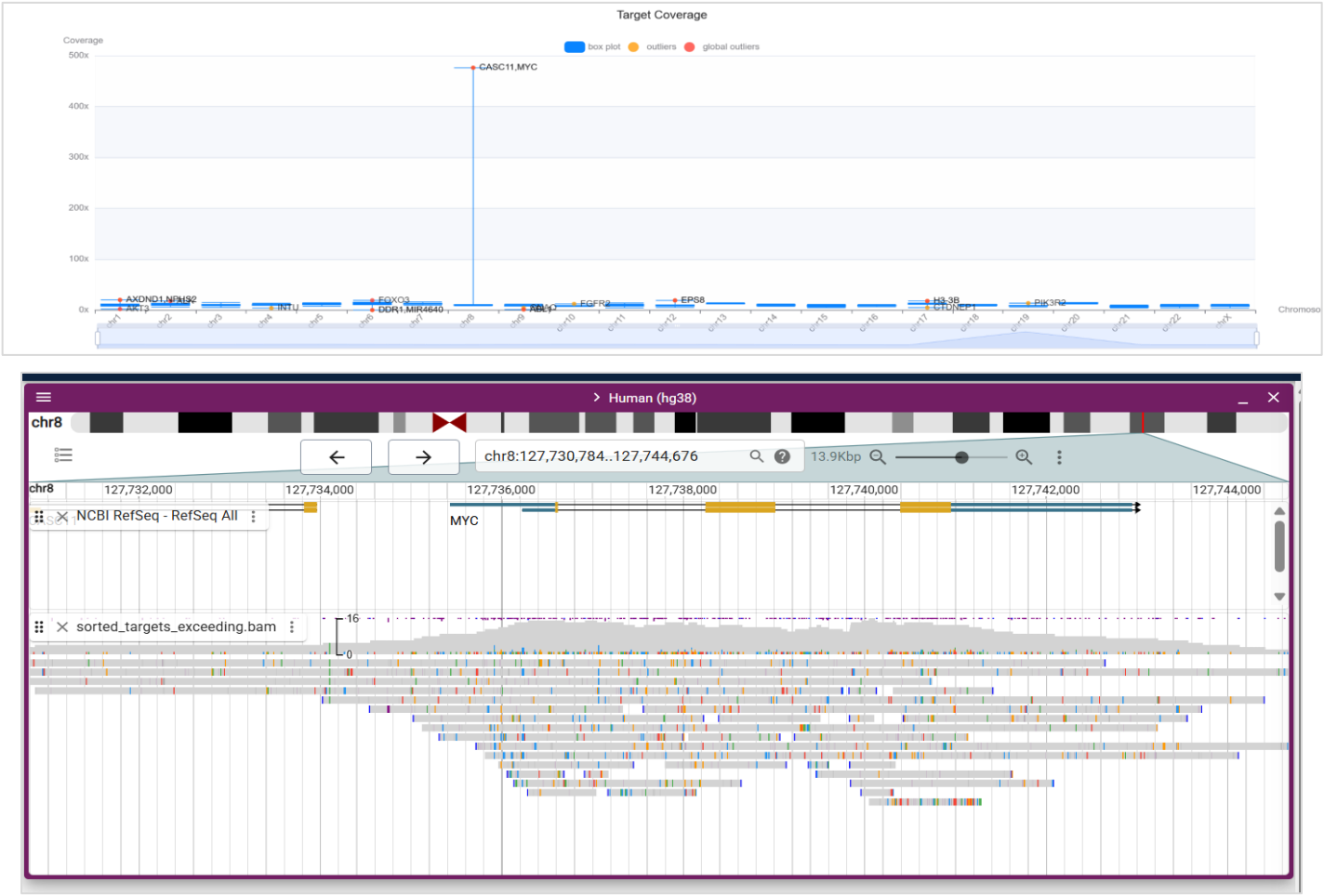


Figure 3.14 Target coverage outliers highlighting MYC on chromosome 8. Top: *Box plot showing target sequencing coverage distribution across all chromosomes. Each box shows the range with median coverage per chromosomes and whiskers shows the range. Red points mark global outliers. MYC on chromosome 8 shows high-coverage outlier.* ***Bottom:*** *Genome browser (JBrowse 2) view of chr8:127,730,784-CHr8:127,744,676 region encompassing MYC. The high read coverage across the gene is consistent with the high coverage outlier detected in the box plot, suggesting amplification of this gene.*

One advantage of ONT long-read sequencing compared to short-read sequencing is the identification of fusions. In this cohort, two cases had clinically relevant fusions identified by standard of care sequencing. Firstly, *KIAA1549::BRAF* fusion in Pilocytic astrocytoma, the ROBIN pipeline revealed one read with the same read ID with high confidence mapping two distinct loci on chromosome 7 which corresponds to *KIAA1549* and *BRAF* and supports a fusion event. Secondly *PSD3::CPNE3* fusion in diffuse astrocytoma MYB or MYBL altered. ROBIN detected two independent long reads which mapped with high confidence and showed split alignment between *PSD3 locus and CPNE3* which supports a fusion event (Figure 3.15). However, fusions are called in ROBIN when there are at least 3 supporting reads and therefore these fusions were not called.

Search

chromBED	BS	BE	Gene	chrom	mS	mE	readID	mapQ	strand	Read Map Start	Read Map End	Secondary	Supplementary	mapping span	tag	Color
chr7	138844464	138867986	KIAA1549	chr7	138857750	138863152	122a571c-157d-4d5a-a255-ddb45ca86a05	60	+	0	5396	false	true	5402	BRAF,KIAA1549	#279f68
chr7	140726418	140924779	BRAF	chr7	140777674	140791057	122a571c-157d-4d5a-a255-ddb45ca86a05	60	+	81	13479	false	false	13383	BRAF,KIAA1549	#279f68

psd3, cpne3

chromBED	BS	BE	Gene	chrom	mS	mE	readID	mapQ	strand	Read Map Start	Read Map End	Secondary	Supplementary	mapping span	tag	Color
chr8	18527303	19084730	PSD3	chr8	18670657	18677830	f92d69c2-2aeb-4212-b2a1-ae37268141ee	60	+	864	8012	false	false	7173	PSD3, CPNE3	#082c61
chr8	18527303	19084730	PSD3	chr8	18670722	18674249	8beeb350-c551-47da-b0e3-1162e1e3edb0	60	+	0	3448	false	true	3527	PSD3, CPNE3	#84909b
chr8	86514435	86561498	CPNE3	chr8	86546304	86547609	8beeb350-c551-47da-b0e3-1162e1e3edb0	60	-	0	1269	false	true	1305	PSD3, CPNE3	#84909b
chr8	86514435	86561498	CPNE3	chr8	86546329	86547145	f92d69c2-2aeb-4212-b2a1-ae37268141ee	60	-	0	816	false	true	816	PSD3, CPNE3	#082c61

Figure 3.15 Evidence supporting gene fusion events. Top: *A single long read aligns to two loci on chromosome 7 corresponding to KIAA1549 and BRAF. The read is split into two high confidence (MAPQ = 60) alignments with one read labelled as supplementary which indicates a breakpoint consisting with KIAA1549::BRAF fusions.* ***Bottom:*** *Two independent reads with high confidence alignments to two loci on chromosome 8; PSD3 and CPNE3 providing supporting evidence of gene fusion.*

3.4.4 Long-read sequencing compared to short-read WGS and cancer panel sequencing

Of the 8 surgical cases sequenced (four retrospective and four prospective), 4 cases have complementary Illumina NGS WGS and 2 have NGS cancer panel sequencing. The results from these two clinical tests were compared to the results generated from ONT long-read sequencing. The type of NGS testing for each case is shown in Table 3.6.

Table 3.6 Overview of standard-of-care next generation sequencing performed for each case. Whole genome sequencing or cancer panel sequencing.

Case ID	Next generation sequencing
SH1343-2020	Whole genome sequencing
SH214-2021	Whole genome sequencing
SH451-2023	Whole genome sequencing
SH576-2023	Whole genome sequencing
SH946-2025	Cancer panel sequencing
SH1049-2025	Cancer panel sequencing

Four cases underwent Illumina whole genome sequencing. For cases SH1343-2020, both short-read and long-read WGS identified variants of clinical significance in the following genes; *IDH1* (c.395 G>A p.Arg132His)), *KRAS* (c.35G>A p.(Gly12Val)), *PIK3CA* (c.1633G>A p.(Glu545Lys)), as well as copy number losses of short arm of chromosome 1 (1p), long arm of chromosome 19 (19q) and short arm of chromosome 9 (9p). Additionally, short-read WGS detected variants in *ARID1A* (c.5548dup p.(Asp1850fs)) and *FUBP1* (c.214del p.(Gln72fs) which were not identified by long-read (Figure 3.16).

For Case SH214-2021 WGS reported a clinically actionable variant of the *BRAF* gene (c.1799T>A p.(Val600Glu) which was not identified by long-read sequencing (Figure 3.17).

For case SH451-2023, both short-read and long-read WGS identified copy number loss of chromosome 8 and *PSD3::CPNE3* fusion (Figure 3.18).

For case SH576-2023 there were several differences between variants detected by short-read and long-read WGS. Short-read sequencing detected deletions of the *DMD* gene and *AMER1,* neither of which were detected by long-read sequencing. Long-read sequencing identified amplification of the following genes: *EGFR, BRAF, MET< FGFR2, FGFR1, MYC, CDK6, IGF2, BIRC7 and ASNS.* Copy number loss of short arm of chromosome 11 (11p) was detected by both short-read and long-read sequencing (Figure 3.19).

Two cases underwent targeted cancer panel sequencing. Variants were assessed in a targeted set of genes relevant to CNS tumour diagnostics, including *ATRX, BRAF, CDKN2A/B, DAXX, EGFR, H3C2, H3C3, H3-3A, H3-3B, IDH1, IDH2, PIK3CA, TERT, TP53 and VHL.* For case SH946-2025, no variants were detected in this set of genes. Amplification in EGFR and MET was also assessed but again no variants were detected. Long-read ONT sequencing similarly revealed no pathogenic variants or amplifications in those genes.

For case SH1049-2025, both targeted panel sequencing and long-read ONT sequencing identified variants predicted clinical significance was detected in the *IHD1* gene (c394C>A p. (Arg132Ser)) in *TP53* (c.817C>T p.(Arg273Cys)). Panel sequencing also identified a second *TP53* variant (c.451C>T p, (Pro151Ser)) which was not identified by long-read sequencing. In addition to the 2 variants detect by LR sequencing, copy number gains of chromosomes 7 and 9 were also observed; these alterations were not assessed by the targeted panel (Figure 3.20).

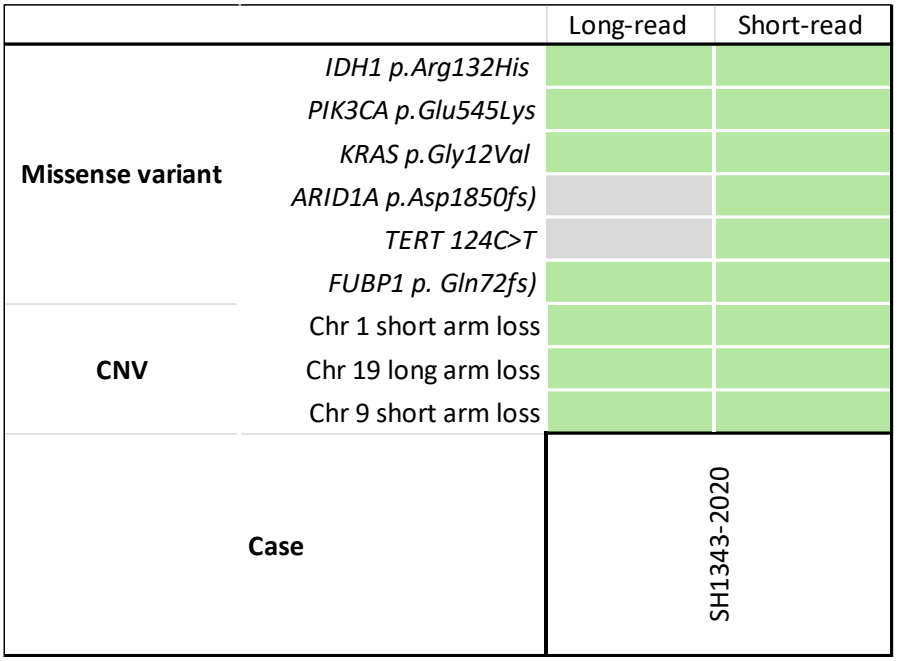


Figure 3.16 Long-read whole genome sequencing results compared to short-read whole genome sequencing results for an Oligodendroglioma (SH1342-2020). *Green shading indicates variant detect, grey shading indicates variant not detected.*

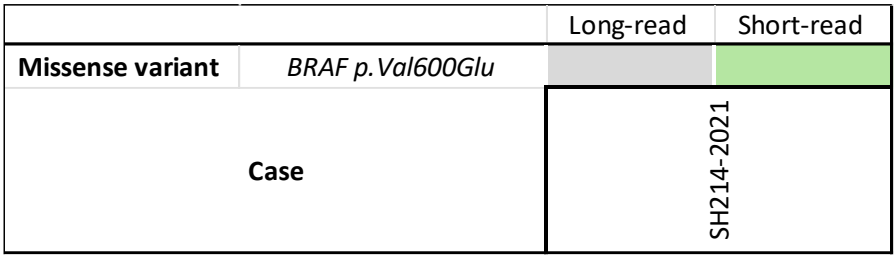


Figure 3.17 Long-read whole genome sequencing results compared to short-read whole genome sequencing results for an Langerhans cell histiocytosis (SH214-2021). *Green shading indicates variant detect, grey shading indicates variant not detected.*

		Long-read	Short-read
Fusion	*PSD3::CPNE3 fusion*		
CNV	Chr 8 loss		
Case		SH451-2023	

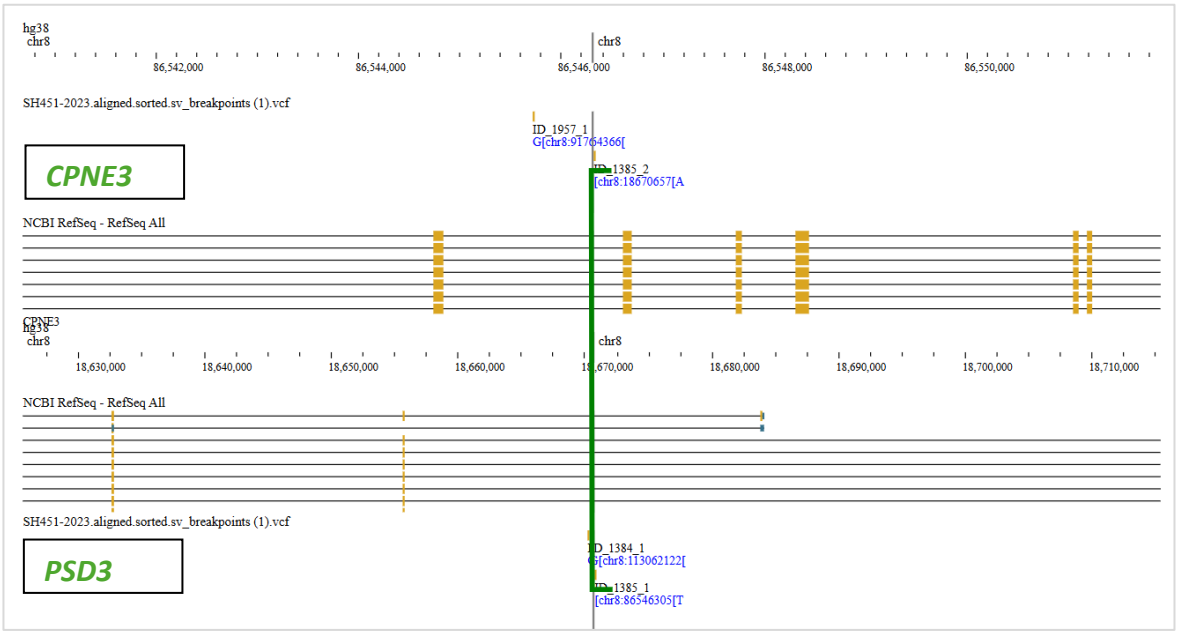


Figure 3.18 Long-read whole genome sequencing results compared to short-read whole genome sequencing results for an Low grade glioma, MYB/MYBL1 altered (SH451-2023). Top: *Shows PSD3::CPNE3 fusion was detected by both short-read and long-read sequencing. Green shading indicates variant present).* ***Bottom:*** *Genome browser view of PSD3::CPNE3 fusion generated by the long-read sequencing data. The top track shows the breakpoint location on the reference sequence of CPNE3 gene loci and the bottom track shows the breakpoint location on the reference sequence of PSD3 gene loci.*

		Long-read	Short-read
Amplification	*EGFR*		
	BRAF		
	MET		
	FGFR2		
	FGFR1		
	MYC		
	CDK6		
	IGF2		
	BIRC7		
	ASNS		
Deletion	*DMD*		
	AMER1		
CNV	LOH Chr 11		
Case		SH576-2023	

Figure 3.19 Long-read whole genome sequencing results compared to short-read whole genome sequencing results for an Mesenchymal tumour – Rhabdosarcoma (SH576-2023). *Green shading indicates variant detect, grey shading indicates variant not detected.*

		Long-read	Short-read
Missense variant	*IDH1 p.Arg132Ser*		
	TP53 p.Arg273Cys		
	TP53 p.Pro151Ser		
CNV	Chr7 gain		
	Chr10 gain		
Case		SH049-2025	

Figure 3.20 Long-read whole genome sequencing results compared to short-read panel sequencing results for an Astrocytoma (SH1049-2025). *Green shading indicates variant detect, grey shading indicates variant not detected.*

3.5 Discussion

In this chapter the performance of ROBIN analytical pipeline, which utilises four methylation based classifiers using both retrospective and prospective samples, was compared against standard of care integrated diagnosis. The results demonstrated that the ROBIN pipeline is successful in classifying cases in a clinically relevant timeframe and can provide accurate classification in most cases but also highlights classifier limitations in diagnostically challenging tumour entities.

From the cohort, four exemplary cases (SH214-2021, SH451-2023, SH576-2023, and SH946-2023) were selected to simulate prospective ONT sequencing with methylation classification. For these cases, the mean number of days from date of biopsy to a final integrated diagnosis to be available was 35 (range: 25-45 days). In addition, cases that received short-read sequencing had a mean turnaround time of 223 days (range 202-257). In comparison, using ONT sequencing and ROBIN pipeline, a methylation based classification was available within 3 hours of sample receipt, and a full integrated diagnosis including copy number variants and MGMT promoter methylation within 24 hours. Furthermore, a fully integrated diagnosis using ONT sequencing can be made on one frozen tissue sample. The equivalent standard of care integrated diagnosis requires tissue samples for multiple different tests which need different optimal sample preservation methods such as FFPE for immunohistochemistry and EPIC array, and fresh frozen tissue for short-read whole genome sequencing. This requires the biopsy to be split which could result in insufficient tissue being available for all tests.

Retrospective ONT methylation classification of post-mortem cases demonstrated complete concordance with standard of care integrated diagnosis with all cases correctly

within 24 hours of sequencing. High inter-model agreement was observed, with concordance of tumour entity by Sturgeon, NanoDX, Pan NanoDx, and the random forest classifier in five of the six cases classified.

These results demonstrate the robustness of ONT base modification sequencing and ROBIN analytical pipeline despite the challenges with working with post-mortem tissue. Structural integrity of DNA is often compromised in post-mortem, causing DNA degradation and fragmentation, which can reduced quality of extracted DNA and pose challenges for downstream analysis (Shabihkhani *et al,* 2014; Zupanic Pajnic*,* 2025). The extent of DNA degradation is based on several variables including the time between death and freezing of the tissue (post-mortem delay) and storage conditions leading to freeze thaw cycles. However, in this cohort, the post-mortem delay for all cases was less than 48 hours, which is considered relatively short. This is notably within the typical maximum retrieval window of 72 hours required by most UK brain banks, a threshold that is frequently exceeded in practice.

Despite these challenges, previous studies have shown that DNA methylation patterns remain stable in post-mortem tissue, even with prolonged post-mortem delay (Barrachinea and Ferrer, 2009; Gross *et al*, 2016; Jarmasz *et al*, 2019). The successful sequencing and classification of post-mortem tissue samples highlights a key advantage of this approach. The availability of extensive archival tissue collections stored for prolonged periods represent a valuable resource for expanding tumour cohorts, enabling classifier training, and facilitating validation studies of rare tumour entities that access to fresh frozen tissue samples is difficult.

In the retrospective cohort of surplus DNA from surgical biopsies, only 50% of cases were correctly classified by all four classifiers. The two cases misclassified were rare tumour entities. One case, SH214-2021 had a clinical diagnosis of Langerhans cell histiocytosis (LCH). The Sturgeon, NanoDX and random forest classifier misclassified this sample as control tissue, however PanNanoDX classifier correctly identified the tumour as LCH. Similarly, the case clinically diagnosed as mesenchymal tumour – rhabdosarcoma was misclassified by sturgeon (Mesenchymal – Ewings sarcoma family with CIC alteration), NanoDX (GBM RTK III) and random forest (MNG – meningioma). It was correctly classified by PanNanoDX as Sarcoma – Rhabdosarcoma like but at low confidence levels. The low confidence scores and misclassification could suggest that this tumour entity is underrepresented in training datasets. Sturgeon and NanoDX do not include Langerhans cell histiocytosis and Rhabdosarcoma as methylation classes supported by these classifiers. This demonstrates the importance of using a pan- cancer model to classify rare tumour entities.

Of the prospective data, 100% of cases were correctly classified by all 4 classifiers. Following a streamlined workflow, sequencing commenced withing 3 hours of receipt of tissue, and methylation data was generated within minutes of beginning sequencing. Correct tumour classification was achieved within five minutes of analysis, demonstrating the potential of this approach to support real time methylation classification.

As the ROBIN analytical pipeline uses four classifiers, there are benefits to using this approach. With multiple classifiers reaching consistent results, even with varying confidence scores, this strengthens confidence in the classifiers and enhances diagnostic accuracy. Furthermore, when the classifiers do not agree, it demonstrates the importance

of neuropathological interpretation of the data and how the classifiers results need to be in concordance with histological data. This could lead towards a two-test diagnostic approach in which a cryo section stained with H and E to determine histological tumour type and tumour percentage, followed by ONT long read sequencing using the ROBIN pipeline for classification and methylation data and long-read WGS pipeline for additional molecular genetics information.

While these classifiers are essential for diagnostics, reference datasets such as the Heidelberg reference datasets undergo new iterations to include updated molecular signatures and novel tumour entities. This highlights a key challenge of machine learning classifiers; they must be updated or retrained to align with the WHO classification systems and re-validation to ensure clinical accuracy.

In this cohort, MGMT promoter methylation status generated from ROBIIN analytical pipeline was compared to standard of care results and showed a high level of confidence with a concordance in 100% of cases where both sets of data were available. Although the number of cases with available SoC data was limited, the results were observed in both methylated and unmethylated cases, supporting the reliability of this approach on clinically relevant MGMT assessment.

For one of the cases in this cohort the EPIC array failed, which could be due to poor sample quality. As the EPIC array is carried out on FFPE tissue, this is known for being poor quality and requires a step to repair DNA prior to bisulphite conversion (Ruijter *et al*, 2015; Simons *et al*, 2025). Bisulphite conversion involves the conversion of unmethylated cytosines to uracil with methylated cytosines remaining unchanged. However, this leads to further DNA damage and low-quality samples fall below usable quality thresholds. In contrast, Oxford

nanopore sequencing can be performed on high quality DNA from fresh frozen tissue and can directly sequence base modifications without the need for bisulphite conversion and avoiding additional DNA damage and potentially improving methylation profiling of poor-quality samples.

MGMT methylation analysis with the ROBIN pipeline requires higher sequencing coverage than the methylation based classifiers (Decon *et al*, 2025) results were still available within 24 hours of sequencing. This offers the advantage of providing MGMT status as part of a single, integrated assay. Collectively, these results support that the ROBIN pipeline is a comprehensive and time-efficient alternative to conventional EPIC array testing in routine clinical practice.

The ROBIN pipeline can reliably detect clinically relevant genomic alterations including copy number variants (CNVs), gene amplifications and deletions with results in concordance with SoC. Copy number profiling was able to accurately detect chromosomal alterations associated with specific tumour types. For example, in one sample, ROBIN identified the deletion of 1p/19q which is a defining feature of oligodendroglioma. Standard of care testing for this case also identified a loss of the short arm of chromosome 9, which was also visible on the ROBIN generated CNV plot. The results show that ROBIN can produce accurate CNV profiles required for tumour classification. In addition, this pipeline was effective at identify deletions and amplifications. *SMARCB1* is a characteristic of ATRT (Biegel *et al,* 1999; Muller *et al,* 2025). In this data set, two case had a clinical diagnosis of ATRT with *SMARBB1* deletions and the ROBIN coverage boxplots demonstrated consistent coverage across most chromosomes but *SMARCB1* showed a reduced coverage and was

identified as a global outlier. Visual inspection of the aligned reads further confirmed reduced coverage across the gene loci, supporting the presence of a deletion.

MYC amplification is a common characteristic of group 3 medulloblastoma (Northcott *et al*, 2012). One case had a diagnosis of Medulloblastoma, group 3 subtype and the coverage boxplot generated by ROBIN demonstrated increased coverage of *MYC* and was identified as a global outlier. Visual inspection using a genome viewer, further supported an amplification with increased coverage across the gene loci.

In the published research study describing ROBIN (Decon *et al*, 2025), the authors describe using readfish (Payne *et al*, 2021) to apply adaptive sampling. Readfish is a software that enables target sequencing by accepting or rejecting DNA molecules during sequencing which in turn allows enrichment of genomic regions of interest. This allowed Decon *et al*, (2025) to achieve on average 30x coverage. In the data set discussed in this chapter, readfish software and adaptive sampling was not applied to sequencing and data generation. This resulted in a lower average coverage of 20x, with some cases with coverage as low as 10x. Two cases in this cohort, each previously had a clinically relevant gene fusion identified; *PSD3::CPNE3* and *KIAA1549::BRAF* fusion. Within the ROBIN pipeline, independent long reads that mapped with high confidence showed split alignments between the genes in these fusions. However, the number of independent reads did not meet the quality threshold applied and therefore the fusions were not called. It is likely that these fusions were not detected due to low coverage, and implementing adaptive sampling could improve fusion calling. However, not applying readfish allowed for whole genome sequencing and following analysis via ROBIN pipeline, data could be analysed by the long-

read ONT pipeline. This pipeline was successful in identifying both fusions which would be displayed using a genome browser which show the breakpoints and both fusion partners.

Finally in this dataset, long-read whole genome sequencing results were compared to short-read sequencing results for standard of care testing. Cases underwent as part of SoC one of two short-read sequencing test; whole genome sequencing or panel sequencing in which a select number of genes were tested. When comparing the results both concordance and discordance was observed. Long-read sequencing was successful in detecting clinically significant single nucleotide variants including *IDH1, PIK3CA, KRAS, DMD, AMER1* and *TP53* and copy number variants including 1p/19q codeletion, loss of chromosome 9p and loss of chromosome 11p. However, some variants were missed by long-read sequencing. *BRAF* p.V600E mutation is a common molecular characteristic of Langerhans cell histiocytosis (LCH) (Shimizu *et al*, 2023). This variant was not successfully detected by the long-read sequencing data. This mutation has a low variant allele frequency (VAF) and therefore would require high sequencing coverage to be detected. The coverage of this case in this cohort was extremely low, 10x, and therefore could not be detected. To be able to move to the approach of combining methylation-based classification with whole genome sequencing, overall sequencing coverage would need to be substantial to detect low frequency variants.

3.5.1 Main Conclusions

- **Standard of care testing compared to ROBIN:** The turnaround time of integrated diagnosis using ONT and ROBIN pipeline is 24hrs which is significantly shorter than the average of 35 days taken for a standard of care integrated diagnosis. It is conceivable that future laboratories concerned with rare cancer diagnosis may move to a totally frozen-section-based workflow (PrestoCHILL) freezing with excellent morphology plus ONT sequencing). This would ensure rapid state-of-the art integrated diagnostics and result in financial savings as iterative testing could be phased out.
- **Methylation classification using ROBIN:** Post-mortem cases show 100% concordance with standard of care diagnosis, with accurate classification within 24 hours of sequencing. This demonstrates the robustness of the four classifiers despite the challenges with working with post-mortem tissue.
- **Pan-cancer classifiers:** This classifier was more reliant when classifying rare tumour entities and were able to correctly classify tumours that the other classifiers were unable to.
- **CNV's, amplifications and deletions:** ROBIN reliably detected CNV's, amplifications and deletions that characterise tumour entities.
- **Adaptive sampling:** Lack of adaptive sampling using readfish resulted in low sequencing coverage and subsequently limited the sensitivity for fusion detection.
- **Long-read vs short-read**: Long-read sequencing results showed concordance with short-read standard of care results for some clinically significant variants. However, low frequency variants such as *BRAF p.V600E,* was not detected indicating the need for higher sequencing depth.

Chapter 4 Molecular pathological architecture of MCS using bulk and single cell sequencing – SR vs LR

4.1 Aims of chapter

Molecular pathological architecture of MCS in human samples is currently unknown as currently studies have been conducted on mouse models or iPSC. One aim of this chapter is to use bulk long-read whole genome sequencing to see if the *HEY1::NCOA2* and its breakpoints be detected. In addition, what are the somatic driver landscape beyond the fusion gene including SNV's, indels and SV and are there any gene specific therapeutic targets. In a subset of cases, we will perform matched short-read whole genome sequencing to compare and answer what can long read sequencing add.

Methylation profiles of cancers aid diagnostics and improve accuracy of diagnosis. They can also be used as prognostic markers and inform effective treatment strategies. An aim of this chapter is to use long-read sequencing (ONT) to investigate the epigenetic landscape of MCS and identify regions of hypermethylation or hypomethylation.

Finally in this chapter we will use long-read single-nuclei sequencing to see which subset of cells in the tumour express the fusion transcript and if the expression is restricted to a particular differentiation state of lineage. Are there particular epigenetic modifications characterising fusion positive versus fusion negative cells? We hypothesize the fusion gene may be epigenetically silenced in the differentiated component of the tumour.

4.2 Introduction

Chondrocytes are the only cells that form cartilage. Mesenchymal stem cells are multipotent cells that can differentiate into connective tissues including chondrocytes; this process is known as chondrogenesis (Goldring, Tsuchimochi and Ijiri, 2006). Normal chondrocyte differentiations which are shown in Figure 4.1 starts with mesenchymal stems cells condensation meaning the cells migrate and form a cluster of cells to initiate differentiation. This leads to the formation of chondroid progenitor cells which will then differentiate into chondrocytes which can subsequently proliferate, undergo hypertrophy or apoptosis(Goldring, Tsuchimochi and Ijiri, 2006; Yang *et al*, 2022). This process of chondrogenesis is controlled by complex interactions of various signalling pathways including TGF-β pathway, Hedgehog pathway and Notch signalling pathway (Yang *et al,* 2022).

HEY1 gene encodes for a transcriptional repressor protein that is a member of basic helix-loop-helix protein family that is a downstream effector of the notch signalling pathway which plays a crucial role in chondrogenesis (Baus, Kabak and Kadesch, 2009; Tanaka *et al,* 2023). The HEY1 has a basic helix-loop-helix domain which has two functional parts; E-box which binds to specific DNA sequences and helix-loop-helix domain which binds to other basic helix-loop-helix proteins to form functional complexes (Jones, 2004). The E-box sequences bind to the promoter regions of target genes which in turn repress the Notch signalling pathway.

NCOA2 is a nuclear receptor coactivator essential for nuclear hormone receptors which are important in regulation of cellular processes such as proliferation, differentiation and cell death. In addition to NCOA2 fusion in MCS, NCOA2 fusions are present in other paediatric

cancer types such as *MOZ::NCOA2* fusion in acute myeloid leukaemia and *PAX1::NCOA2* in rhabdomyosarcoma (Sankhe, Hall, and Kendall, 2025).

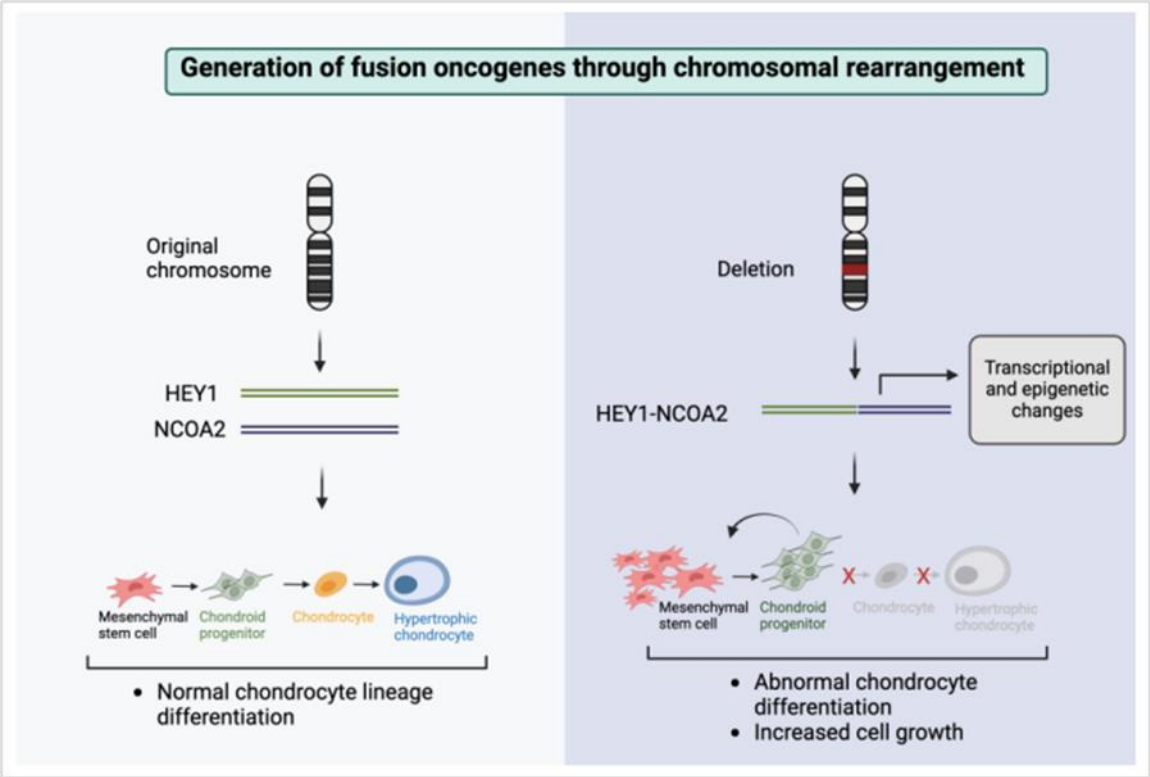


Figure 4.1 Effect of HEY1:NCOA2 fusion gene of cell proliferation in mesenchymal chondrosarcoma. Left: *In the normal chromosome, the HEY1 and NCOA2 genes remain separate, allowing transcriptional regulation and normal chondrocytes lineage differentiation from mesenchymal stem cells through to chondrocytes.* ***Right:*** *A 10 Mb deletion causes the HEY1::NCOA2 fusion gene resulting in altered transcriptional regulation. The fusions disrupt normal chondrocytes differentiation, promoting abnormal cell growth (Figure courtesy of Casmir Turnquist, 2025).*

HEY1::NCOA2 fusion results from chromosomal rearrangement of the DNA-binding domain of *HEY1* (NOTCH pathway transcription signalling repressor) with transcriptional activation domains of NCOA2. This converts *HEY1* into an oncogenic transcription activator that activates NOTCH target genes leading to impaired cellular differentiation and disrupted maturation (Wang *et al,* 2012; Panagopoulos *et al*, 2014; Tanka *et al*, 2023).

Recent genomic studies using iPSC, have demonstrated that the fusion acts as an oncogenic transcription factor by altering HEY1 dependant gene regulation (Qi *et al,* 2022). The fusion protein binds HEY1 target promotors but converts transcriptional repression into activation which results in upregulation of genes involved in cell cycle progression, survival, and development of signalling pathways, including PDGF/PI3K-AKT, WNT, and Hedgehog signalling, as well as genes associated with chondrocyte differentiation (Qi *et al,* 2022).

Currently, there are no targeted treatment options for MCS patients, and as conventional chemotherapy and radiotherapy do not substantially improve survival, these targeted treatment options are urgently needed. Currently only mouse models or *in vitro* studies have been performed to examine the functions of the *HEY1::NOCA2* fusion gene. There is no human whole genome sequencing data set for this cancer type.

4.3 Methodology

Mesenchymal chondrosarcoma samples were received from University of Leiden, Department of Pathology. These cases were used under ethical approval from London – Stanmore Research Ethics Committee (REC 17786; IRAS 45163).

Eleven cases with frozen tissue available from surgical biopsies were selected for Short-read and Long-read whole genome sequencing (bulk and single cell) and Case demographics are described in Table 3-1. All tumours were from the primary site, although the site of the tumour varied: Leg (3), Hand (1), Thigh (1), Scapula (2), Pelvis (2), Sacrum (1), femur (1).

Table 4.1 Case demographics. *All tumour samples were from the primary site. Short-read WGS was performed on 10 cases, Long-read WGS on 9 cases and LR Single-cell on 3 cases.*

Sample ID	Age at diagnosis	Tumour group	Anatomical Site	Molecular analysis to confirm fusion	SR-WGS	LR-WGS	LR-SC
17R1075	34	Primary	Leg	Yes - RT-PCR	Yes	No (failed)	No
18R3548	24	Primary	Leg	Yes - RT-PCR	Yes	Yes	No
2013-1070	17	Primary	Hand	Yes - RT-PCR	Yes	Yes	No
020054	70	Primary	Thigh	Yes - RT-PCR	Yes	Yes	No
010156	27	Primary	Leg	Not done	Yes	Yes	Yes
980479	94	Primary	Scapula	Not done	Yes	Yes	No
18R3967	56	Primary	Scapula	Yes - RT-PCR	No	Yes	No
17R0067	40	Primary	Pelvis	Yes - RT-PCR	Yes	Yes	No
23R2886	48	Primary	Pelvis	Yes - RT-PCR	Yes	Yes	Yes
940743	23	Primary	Sacrum	Not done	Yes	Yes	Yes

To assess the quality of tissue and tumour content, 10 µm section from cryo blocks were stained with Haematoxylin and Eosin (H&E) using an automated Shandon Linistain GLX Random Access Stainer (ThermoFisher Scientific, USA).

4.3.1 DNA extraction using Monarch HMW DNA extraction kit for Tissue and quality control

DNA was extracted from x30µm cryo sections using the Monarch HMW DNA Extraction Kit for Tissue (New England Biolabs, England) following the manufactures protocol. DNA was eluted in 100µl EB buffer. DNA concentration was measured using the Qubit 4 Fluorometer (ThermoFisher Scientific, USA). DNA samples were prepared using the QubitTM dsDNA BR Assay Kit. The purity of DNA was measured using the Nanodrop 2000/2000c Spectrometer (ThermoFisher Scientific, USA). DNA fragment size was measured using Agilent 4200 TapeStation System (Agilent Technologies, USA). Samples were prepared using Agilent Genomic DNA Screen Tape and Reagents following manufacturing protocols.

DNA samples were sheared using Covaris g-TUBE (SKU-520079, Covaris). The DNA sample were transferred to g-TUBE and centrifuged at 7200 RPM for 60 seconds. Following this the g-TUBE was inverted and centrifuged under to same conditions. This was for a targeted DNA size of 20kbp.

4.3.2 Short-Read Whole Genome Sequencing

DNA libraries were prepared and sequenced by GENEWIZ (Oxford, UK) using Illumina® NovaSeq™ platform. DNA was sequenced to an average depth of 40X. FASTQ files were returned for downstream bioinformatic analysis.

4.3.3 Short-Read Whole Genome Sequencing Analysis

Bioinformatic analysis was performed by Dr Ebony Cave using the following pipeline. Oncoanalyser Nextflow Pipeline version 2 which was developed by Hartwig Medical Foundation, Australia (Ewels *et al,* 2020) was used for the whole genome sequencing analysis. Sequence reads were aligned to Human GRCh38 reference genome using BWA-MEM2. Single-nucleotide variant (SNV), multi-nucleotide variant (MNV) and Indell calling were performed using SAGE and PAVE. Structural variant (SV) calling was performed using ESVEE and LINX. Copy number variants (CNV) were detected using AMBER COBALT PURPLE. Clinical reports were generated using orange. See below for full code:

```
nextflow run nf-core/oncoanalyser \
-config refdata.local.config \
-revision 2.0.0 \
-profile singularity \
--mode wgts \
--genome GRCh38_hmf \
--input spreadsheet_for_oncoanalyser.csv \
--outdir ${oncoanalyser_output_folder_name}
```

The ‘-config’ file and ‘—input’ spreadsheet can be found in appendix 3.

4.3.4 Long-Read ONT Library prep and Whole Genome Sequencing

Samples were prepared with the ONT Ligation sequencing kit V14 with an adjusted protocol. 2 µg DNA was repaired and prepared for adapter ligation purified using AmpureXP DNA binding beads (Beckman Coulter; USA). Library concentration was measured using Qubit™ 4 Fluorometer and Qubit™ dsDNA HS Assay kit prior to flow cell loading. The library was loaded onto R10.4.1 PromethION flowcell (FLO-PRO114M, Oxford Nanopore Technologies, UK). Basecalling was performed whilst sequencing using Dorado (version 7.6.7) which is integrated within MinKNOW (version 24.11.8). Reads were called using High-accuracy (HAC) model (v4.3.0, 400bps) with 5hmC and 5mC modifications.

4.3.5 Long-Read Whole Genome Sequencing Analysis

Long-read ONT Pipeline, which was developed by Andrew Beggs, University of Birmingham (2024) was used for the whole genome sequencing analysis as described in section 3.3.5 and performed by Dr Ebony Cave.

4.3.6 Single-nuclei isolation and GEM-X formation

Nuclei isolations were prepared according to the 10x Chromium Nuclei Isolation Kit (10X Genomics, USA) protocol with all steps performed on ice. Frozen MCS tissue (~50 mg) was added to a prechilled sample dissociation tube (10X Genomics) that had be coated in 10% BSA overnight (Sigma-Aldrich, UK). 200 µl Lysis Buffer (Lysis Reagent (2000558, 10X Genomics), Reducing Agent B (2000087, 10X Genomics), Surfactant A (2000559, 10X Genomics) was added and the sample dissociated for 20 seconds using the pestle provided in the kit. A further 300 µl Lysis Buffer was added and incubated on ice for 10 minutes with dissociation with the pestle every 2 minutes to ensure full dissociation of the tissue. The dissociated tissue was transferred to a pre-chilled Nuclei Isolation Column (10X genomics) and centrifuged at 16,000 rcf for 20 seconds at 4 °c. The flowthrough was transferred to a BSA coated Eppendorf tube and centrifuged at 500 rcf for 3 minutes at 4 °c. The supernatant was removed the pellet resuspended in 700 µl Debris Removal Buffer (Debris Removal Reagent (2000560, 10X Genomics), Reducing Agent B (2000087, 10x Genomics) and subsequently centrifuged at 700 rcf for 10 minutes at 4 °c. The supernatant was discarded and the pellet resuspended in 700 µl Wash and Resuspension Buffer with RNAase Inhibitor (1x PBS, 10% BSA, RNase Inhibitor (200565, 10X Genomics) and centrifuged at 500 rcf for 5 minutes at 4 °c. This step was repeated twice to remove debris. The supernatant was removed and nuclei resuspended in 200 µl Wash and Resuspension Buffer by gentle pipetting. Nuclei were filtered using 40 µm Flowmi (Sigma-Aldrich, UK) filter tips, this step was repeated using 100 µl Wash and Resuspension Buffer to flush filters. Nuclei were counted using AO/PI Luna Fx7 (Labtech, UK) by adding 9 µl of sample to 1 µl of dye.

GEM generation and barcoding was performed according to the 10X Chromium GEM-X Single Cell 5’ Kit protocol. These steps were performed by Andrew Lee. GEM-X chip was

prepared by loading nuclei with Master Mix (RT Regent E (2001106, 10X Genomics), Poly-dT RT Primer B (2001110, 10X Genomics), Reducing Agent B (2000087, 10X Genomics), RT Enzyme E (2001105, 10X Genomics)), GEM-X Single Cell 5' Gel Bead v3 (2001129, 10X Genomics) and Partitioning Oil B (2001213, 10X Genomic). Chromium X was run for GEM generation using firmware version 2.4.0 Following this 100 µl GEM were transferred from the recovery wells into a tube strip on ice. GEMs were incubated on a thermal cycler using the following GEM-RT Incubation protocol:

Table 4.2 GEM-RT incubation protocol (Adapted from Chromium GEM-X Single Cell 5' Reagent Kits v3, 10X Genomics, 2024)

Lid Temperature	**Reaction volume**	**Run time**
48°c	125 µl	55 minutes
Step	**Temperature**	**Time**
1	48°c	45 minutes
2	85°c	5 minutes
3	4°c	Hold

Post GEM-RT Incubation, 125 µl Recovery Agent (220016, 10X Genomics) was added to the sample and incubated at room temperature for 2 minutes. 125 µl Recovery Agent and Partitioning Oil was removed and 200 µl of Dynabeads Cleanup Mix (Cleanup Buffer (2000088, 10X Genomics), Dynabeads MyOne Silane (2000048, 10X Genomics), Reducing Agent B (2000087, 10X Genomics), Nuclease-free Water) was added to the tube and incubated for 10 minutes at room temperature. The tubes were placed on 10X magnetic separator (high position) until the solution clears. The supernatant was removed and pellet washed with 300 µl 80% ethanol for 30 seconds. Ethanol was removed and a further 200 µl

was added and removed. The pellet was resuspended in 35.5 µl Elution Solution I and incubated for 1 minute at room temperature. This was placed back on the 10X magnet separator (low position) until the solution clears and 35 µl of sample removed. 65 µl of cDNA Amplification Reaction Mix (Amp Mix (2000047, 10X Genomics), cDNA Primers (2000089, 10X Genomics)) was added for cDNA Amplification and the sample incubated in the thermal cycler using the following protocol:

Table 4.3 Sample incubation protocol for cDNA amplification (Adapted from Chromium GEM-X Single Cell 5' Reagent Kits v3, 10X Genomics, 2024)

Lid Temperature	**Reaction Volume**	**Run Time**
105°c	100 µl	~ 40 minutes
Step	**Temperature**	**Time**
1	98°c	45 seconds
2	98°c	20 seconds
3	63°c	30 seconds
4	72°c	1 minute
Repeat steps 1-4 for 12 cycles (target cell recovery > 6000)		
6	72°c	1 minute
7	4°c	Hold

Following cDNA amplification 60 µl Beckman Coulter™ SPRIselect™ reagent (15605838, Fisher Scientific) is added and incubated for 5 minutes at room temperature. The sample tubes are placed on 10X magnetic separator (high position), supernatant removed and pellet washed with 200 µl 80% ethanol. The ethanol is removed and this step repeated. The

pellet is resuspended in 40.5 µl Buffer EB and returned to the magnet (low position). 40 µl sample is removed and retained for QC steps.

cDNA fragment size was measured using Agilent 4200 TapeStation System (Agilent Technologies, USA). Samples were prepared using Agilent D5000 High Sensitivity Screen Tape (5067-5592, Agilent) and Reagents (5067-5593, Agilent) following manufacturing protocols.

4.3.6 ONT Library Prep for Single-nuclei Whole Genome Sequencing

ONT library preparation was carried out using ONT Ligation Sequencing kit (SQK-LSK114) following the Single-cell transcriptomics sequencing from 5' cDNA prepared with 10x Genomics using SQK-LSK114 protocol (version SST_9204_v114_revL_09june2025) (Oxford Nanopore Technologies, UK). For PCR amplification, 10 ng of cDNA amplicons produced in step 3.3.6 were incubated with the custom oligo sequences described below in Table 4.4 in a thermal cycler using the following conditions:

Table 4.4 ONT PCR amplification of cDNA amplicons thermal cycler protocol *(adapted from Single-cell transcriptomics sequencing from 5' cDNA prepared with 10x Genomics using SQK-LSK114, Oxford Nanopore Technologies, 2025)*

Cycle step	Temperature	Ramp rate	Time	No. of cycle
Initial denaturation	94°c	Max	3 minutes	1
Denaturation	94°c	Max	30 seconds	8
Annealing ramp-down	66°c down to 58°c	0.2°c/s	40 seconds	
Annealing	58°c	max	50 seconds	
Extension	65°c	Max	6 minutes	
Final extension	65°c	Max	10 minutes	1
Hold	4°c	-	-	-

40 µl of AmpureXP DNA binding beads (Beckman Coulter; USA) was used for clean-up following PCR amplification. Assuming an average size of 1 kb, 200 fmol of cDNA was carried forward for end-prep and adapter ligation. The sample was purified using purified using AmpureXP DNA binding beads (Beckman Coulter; USA). Library concentration was measured using Qubit™ 4 Fluorometer and Qubit™ dsDNA HS Assay kit prior to flow cell loading. The final library was prepared assuming fragment length of 1 kb with 33ng of library in 32 µl of EB buffer. The library was loaded onto R10.4.1 PromethION flowcell (FLO-PRO114M, Oxford Nanopore Technologies, UK). Sequencing was performed for 72hrs.

Basecalling was performed whilst sequencing using Dorado (version 7.6.7) which is integrated within MinKNOW (version 24.11.8). Reads were called using High-accuracy (HAC) model (v4.3.0, 400bps) with 5hmC and 5mC modifications.

Table 4.5 Custom oligo sequences for PCR amplification in ONT Single cell transcriptomics sequencing from 5' cDNA prepared with 10X Genomics using SQK-LSK114 protocol.

Fwd_3580_partial_read1_defined_for_5'_cDNA	5'-/5phos/ACTTGCCTGTCGCTCTATCTTCCTACACGA
Rev_PR2_partial_TSO_defined_5'_cDNA	5'-/5phos/TTTCTGTTGGTGCTGATATTGCAAGCAGTGGTATCAACGCAGAG-3'

4.3.6 Single-nuclei Whole Genome Sequencing Analysis

Bioinformatic analysis was performed by Dr Ebony Cave. Epi2me-labs/wf-single-cell nextflow workflow version 3.3.2 (Epi2me-labs, 2021, https://github.com/Epi2me-labs/wf-single-cell) which was developed by Oxford Nanopore Technologies, Oxford was used for the single-nuclei whole genome sequencing analysis. Sequence reads were aligned to Human GRCh38 reference genome. An expected cell count of 10,000 cells was used and a minimum read quality of 14 – full length only. Code line for this workflow is shown below:

```
nextflow run epi2me-labs/wf-single-cell --bam
${input_bam_file} --kit '5prime:v2' --ref_genome_dir
HUMAN_GRCh38_REFERENCE_GENOMES/refdata-gex-GRCh38-2024-A --
expected_cells 10000 --min_read_qual 14 --full_length_only
false --gene_assigns_minqv 5 --matrix_min_genes 10 --
matrix_min_cells 1 --matrix_max_mito 50 --call_variants true
-resume -profile singularity
```

Seurat (version 5.3.0) (https://satijalab.org/seurat/) was used for quality control, fusion cell annotation and cell clustering for UMAP generation. Cell identities were assigned by applying label transfer from the two reference data sets: Human Limb Embryo Atlas (atlas1) (Zhang *et al*, 2023) and Endochondral Ossification atlas (atlas 2) (Lawrence *et al,* 2025). Full code can be found in appendix 5.

4.4 Results

4.4.1 Quality control of frozen tissue samples

Of the 9 patients, 6 cases had previously had *HEY1::NOCA2* fusion confirmed by RT-PCR. H&E was used to assess the pathology of the samples, tumour content and viability of the MCS tumour samples (see Table 4.6). H&E showed that all cases had the characteristic areas of primitive small round blue cells and areas of differentiated mature cartilage cells at varying proportions. Tumour content percentage ranged between 60-90% and viability score ranged from 55-100%.

Table 4.6 Pathology of MCS tissue samples

Case ID	Tumour content (%)	Viability (%)	Macroscopic description
17R1075	90	90	
18R3548	70	90	
2013-1070	75	95	
020054	90	90	All small round blue cells
010156	95	100	75% tumour small cell, 25% presumed differentiation to mature cartilage
980479	90	55	
18R3967	70	95	
17R0067	60	95	Mostly small round blue cell
23R2886	-	-	60% small blue cells, 40% differentiation to mature cartilage
940743	70	80	

Prior to sequencing, DNA purity was assessed using Nanodrop 260/280 and 260/230 ratio. 260/280 absorbance ratio ranged between 1.8 and 2.0 which is considered pure DNA. 260/230 absorbance ratios were >2.0 which suggests the samples were not contaminated with salts and other contaminants from the extraction process. These absorbance ratios are indicative of good quality DNA for both short-read and long-read sequencing.

DNA fragment size was measured using Tapestation Genomic DNA screen tape and reagents. Fragment size for all DNA samples was >25 kbp and subsequently samples were sheared using g-TUBE for a fragment size of 20 kbp. Electropherograms showed one distinct upper peak with no 'noise' which is indicative of no fragmentation within the sample. An example of Nanodrop spectrum graph and Tapestation electropherograms are shown in Figure 4.2.

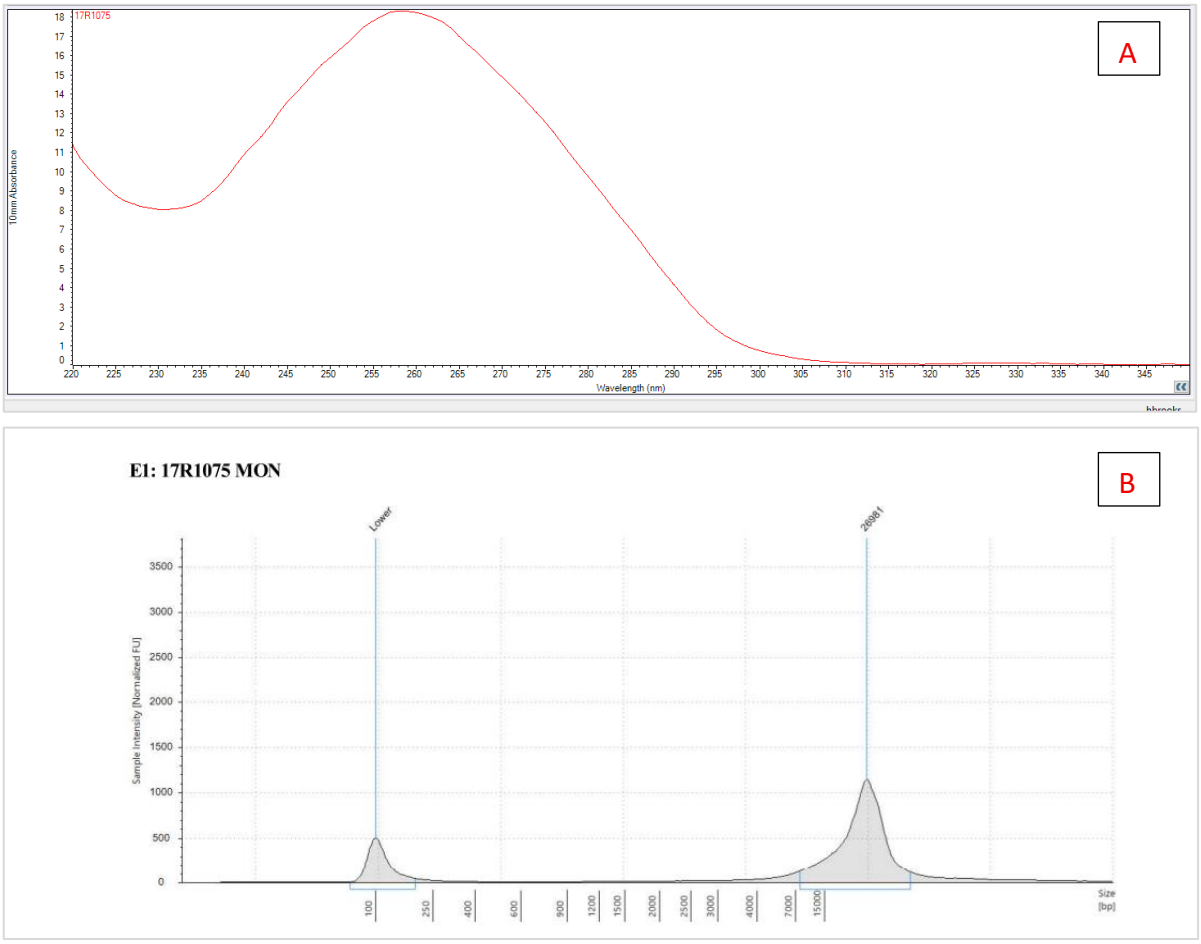


Figure 4.2 Example of Nanodrop and Tapestation report from 17R1075. (A) *Nanodrop full spectrum graph. The broad peak at 260nm indicates the presence of nucleic acids. The 260/280 ratio can be calculated and is indicative of contamination by protein. This sample (17R1075) has a 260/280 ratio of 1.87 which falls into the range of good quality pure DNA sample. 260/230 ratio reflects how pure the sample is from other contaminants such as salt. Optimal range is 2.0 or above. This sample has a 260/230 ratio absorbance of 2.27.* ***(B)*** *Tape station electropherogram shows the size distribution across the sample. A lower marker is included to align with the ladder data. There is a clear distinct peak at 26 kbp and no unexpected peaks which suggests no fragmentation of the sample.*

4.4.2 Short-read vs Long-read sequencing summary

Median sequencing depth coverage for SR WGS was 41X (range 37X-48X). In this cohort, SR WGS detected on average 6 somatic variants (range 1-12), Somatic SNV's 2 (range 0-13). The tumour mutational burden per Mb was high with an average of 12 (range 10.2-17.9). Microsatellite indels per Mb ranged between 0.4-0.8 which is considered stable. Tumour purity was estimated and ranged from 0.16-0.93 and tumour ploidy ranged between 2-2.3.

For LR-WGS cases were sequenced on a promethION flowcell for 72hrs. Median sequencing depth coverage was 18X (range 13X – 30X). In this cohort, LR WGS detected on average 246 variants (range 122-387), 114 coding variants (range 61-191), SNV's 236 (range 120-371) and Indels 10 (range 2-17). The tumour mutational burden (coding and silent variants / Mb) was low with an average of 4.9 (range 2.56-8). Microsatellite indels per Mb ranged between 0.03 -0.32 which is considered stable. Tumour purity was estimated and ranged from 0.6-1 and tumour ploidy ranged between 2-2.15.

4.4.3 Detection of *HEY1::NCOA2* fusion

Short-read and Long-read sequencing both were successful in detecting the *HEY1::NCOA2* fusion. Both sequencing platforms were able to detect the fusion in 88% (n=8) of the cohort. The sample in which the fusion was not detected, (case ID 980479) had not previously had molecular analysis to confirm the fusion. Figure 4.3 shows the breakpoints of the fusion for each sample. As expected, the breakpoints on *HEY1 gene* occurred between exon 4 and 5 (coordinates range chr8:79765758 – chr8:79766612) with exon 5 being deleted. The breakpoints on *NCOA2* for all cases occurred between exon 12 and exon 13 (coordinates range chr8:70142867 – chr8:70147716) with exons 1-12 of the gene being deleted causing the fusion between exon 13 of *NCOA2* to exon 4 of *HEY1.* One sample (23R2886) had two breakpoints identified by LR, first at coordinates chr8:79765922 // chr8:70145965 which falls into the expected fusion occurring between exon 13 of *NCOA2* and exon 4 of *HEY1*. The second break point (chr8:79764509// chr8:70146039) occurs within exon 5 of *HEY1.* Only one break point for this sample at chr8:79765922 // chr8:70145965 was detected in the SR data. The break points for each case only varied by maximum 2 bp when comparing SR versus LR data. See Table 4.7 for breakpoints coordinates.

One case within the cohort did not have the *HEY1::NCOA2* fusion identified by both sequencing platforms but a fusion between *EWSR1::NFATC2* was detected (Figure 4.4). This suggests that this case is not MCS but a different sarcoma entity.

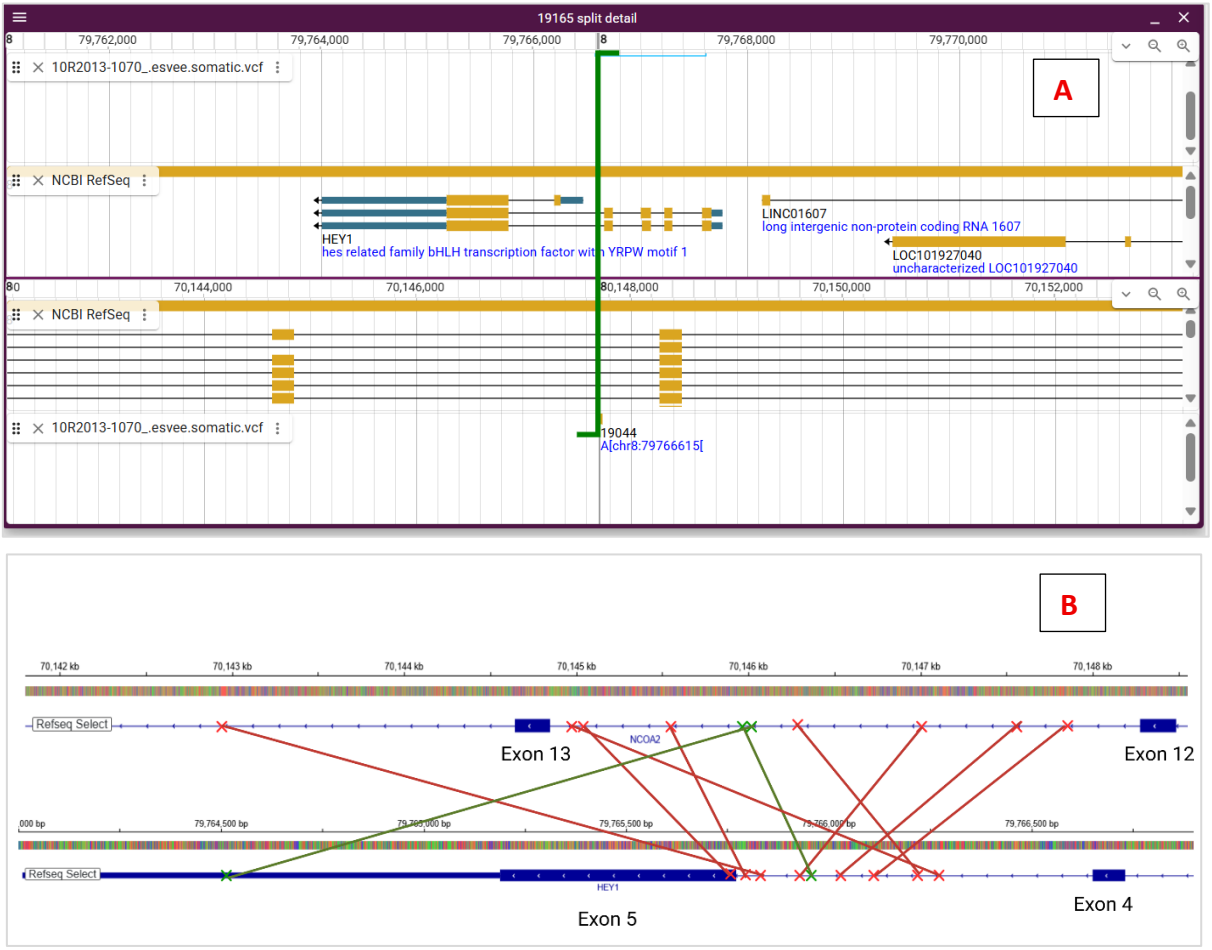


Figure 4.3 Location of HEY1::NCOA2 fusion breakpoints within each sample. *Red X marks HEY1::NCOA fusion breakpoints in each sample with 1 breakpoint identified per sample. Green X marks HEY1::NCOA2 fusion breakpoints in sample (23R2886) with 2 breakpoints identified by LR sequencing.*

Table 4.7 Showing HEY1 and NCOA2 breakpoints for SR and LR data.

Case ID	HEY1::NCOA2 fusion breakpoints (SVIM) LR	HEY1::NCOA2 fusion breakpoints (SVIM) SR
18R003548	chr8:79766265 // chr8:70144965	chr8:79766265 // chr8:70144965
IOR2013-001070	chr8:79766612 // chr8:70147716	chr8:79766614 // chr8:70147715
020054I	chr8:79766200 // chr8:70146225	chr8:79766200 // chr8:70146226
010156I	chr8:79765758 // chr8:70145263	chr8:79765758 // chr8:70145265
980479IOR	No fusion detected	No fusion detected
18R003967	chr8:79765888 // chr8:70146817	NOT SEQUENCED
17R000067	chr8:79765800 // chr8:70145577	chr8:79765800 // chr8:70145577
23R002886	chr8:79765922 // chr8:70145965 chr8:79764509// chr8:70146039	chr8:79765922 // chr8:70145965
940743IOR	chr8:79765922 // chr8:70145965	chr8:79765930 // chr8:70145970
17R1075	NOT SEQUENCED	chr8:79766393 // chr8:70147293

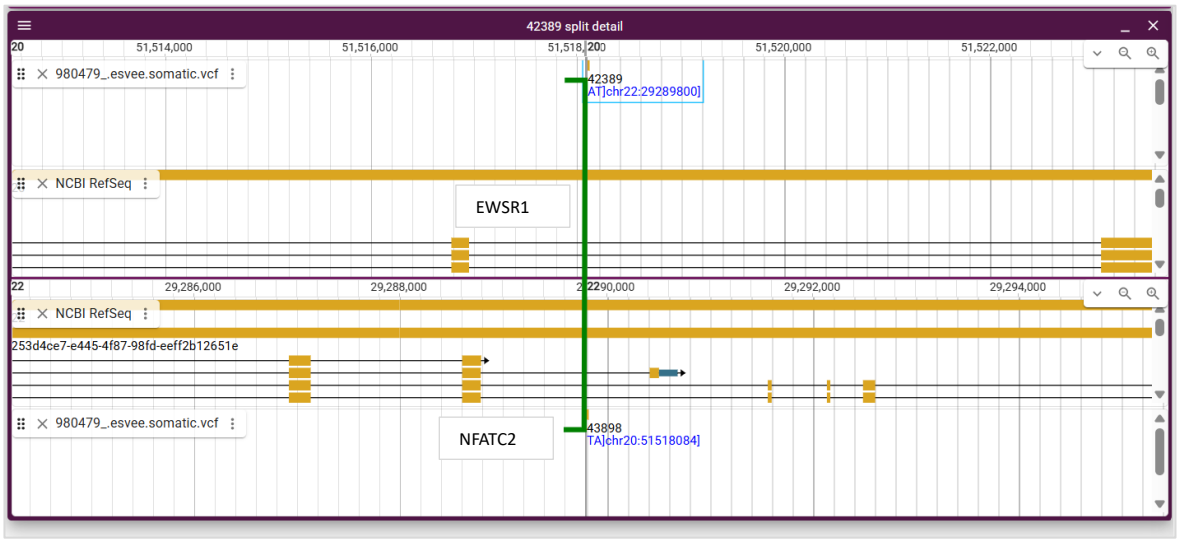


Figure 4.4 JBrowse2 output displaying EWSR1::NFACT2 fusion in sample 980479. *The green line shows the breakpoints and resulting gene fusion.*

4.4.4 Short-read and Long-Read Somatic Variants

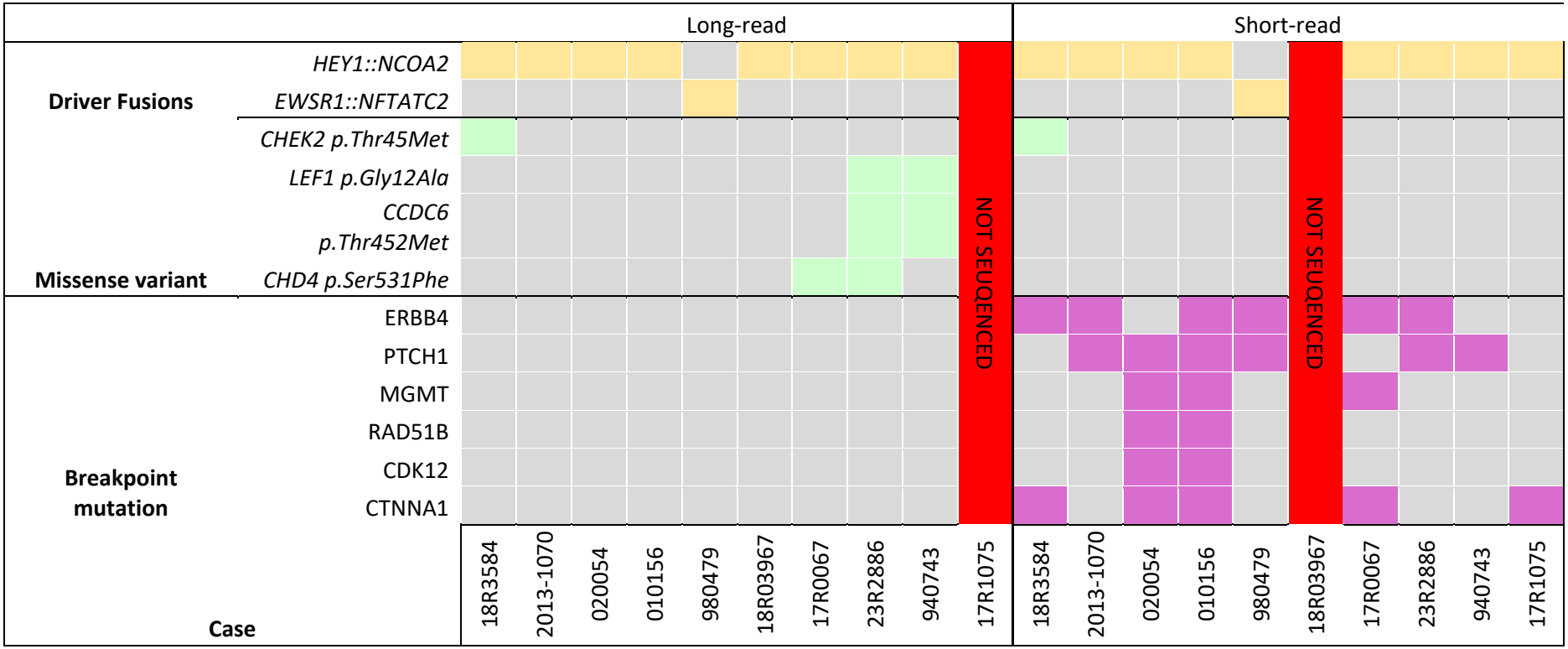
As discussed in the previous section, *HEY1::NCOA2* fusion was detected in 88% of cases by both short-read and long-read sequencing. *EWSR1::NFTATC2* fusion was also detected by both sequencing platforms (n=1, 11%). The most common alterations were breakpoint mutations in ERBB4 which was found in in 66% (n=6) of case by short-read sequencing.

Missense variant in CHEK2 (n=1, 11%) was detected by long-read and short-read sequencing. Other missense variants detected by long-read only in the following genes LEF1 (n=2, 22%), CCDC6 (n=2, 22%), CHD4 (n=2, 22%).

Breakpoint mutations in PTCH1 was detected by short-read (n=6, 66%), MGMT short-read (n=3, 33%), RAD51B by short-read (n=2, 22%), CDK12 by short-read (n=2, 22%), and CTNNA1 by short-read (n=5, 55%).

Table 4.8 shows comparison of somatic variants identified by short-read and long-read whole genome sequencing of 9 MCS cases. A full list of variants detected is in Figure A3 of appendix 8.

Table 4.8 Comparison of genetic somatic variants detected by long-read and short-read sequencing. *Yellow shading represents fusion detected, green shading represents missense variant detected, purple shading represents breakpoint mutation detected, and grey shading represents no genomic event detected.*



4.4.5 Long-ready epigenetic modifications (Methylation)

Base modification was assessed across the *HEY1* locus and 3 Kb upstream. Base modifications of both 5-methylcytosine (5mC) and 5-hydroxmethylcytosine (5hmC) were detected at the *HEY1* locus. A threshold was applied in which more than 80% of the reads had to contain methylation in order for a base to be considered modified. At the time of sequencing patient matched control samples were not available and therefore base modification data was compared to Oxford Nanopore Genome in a Bottle reference sample.

Assessment of base modification up to 3 Kb upstream of *HEY1* showed a similar pattern of methylation when compared to the genome in a bottle reference sample between coordinates Chr8: 79,770,443 and CHr88: 79,770,914. Across the *HEY1* locus, no pattern of methylation was observed.

Figure 4.5 Methylation pattern of CpG sites up to 3Kb upstream of HEY1.

4.4.6 Long-read single nuclei sequencing Library prep QC

Sample 1 (23R2886a)

Following nuclei isolation, nuclei were marked with Acridine orange/propidium iodide (AO/PI) and counted with Luna Fx7 to assess concentration and viability. The concentration of nuclei was 1.27 x10^7 cells/ and the cell viability was 0. As we are isolating nuclei, we are aiming for cell viability to be <5% as any higher would indicate intact cells and failure of nuclei dissociation. For this sample, 40,000 nuclei were loaded onto the GEM-X chip with an expected capture of approx. 80% of nuclei resulting in aimed recovery of 32,000 nuclei. Tapestation was used to measure fragment size and cDNA concentration following cDNA amplification. As expected, a peak was seen on the tapestation electropherogram (Figure 4.6) between 400-2000 bp, with a concentration of 4760 pg/µl. A smaller peak was observed at ≈ 250 bp, however the concentration in comparison to the main peak was low (365 pg/µl). The overall concentration of the sample was 5340 pg/µl, 213ng. 10ng of cDNA was input for ONT library preparation. Table 4.9 shows results of QC steps through ONT library preparation. A final yield of 158 ng was quantified at the end of library preparation which allowed for the recommendation of 100 fmol (1kb, 62 ng) of library to be sequenced.

Table 4.9 QC metrics from ONT library prep. *Average fragment size length of 1 kb was assumed to calculate library to carry forward to next steps.*

Sample	Protocol step	Concentration	Volume	Yield
Sample 1	**PCR amplification**	116 ng/ µl	14 µl	1624 ng
	End prep	9.28 ng/ µl	60 µl	556 ng
	Adapter ligation	4.66 ng/ µl	34 µl	158 ng
Sample 2	**PCR amplification**	55 ng/ µl	14 µl	770 ng
	End prep	1.6 ng/ µl	60 µl	96 ng
	Adapter ligation	4.22 ng/ µl	34 µl	143 ng
Sample 3	**PCR amplification**	61.4 ng/ µl	14 µl	859 ng
	End prep	6.58 ng/ µl	60 µl	394 ng
	Adapter ligation	3.79 ng/ µl	34 µl	128 ng

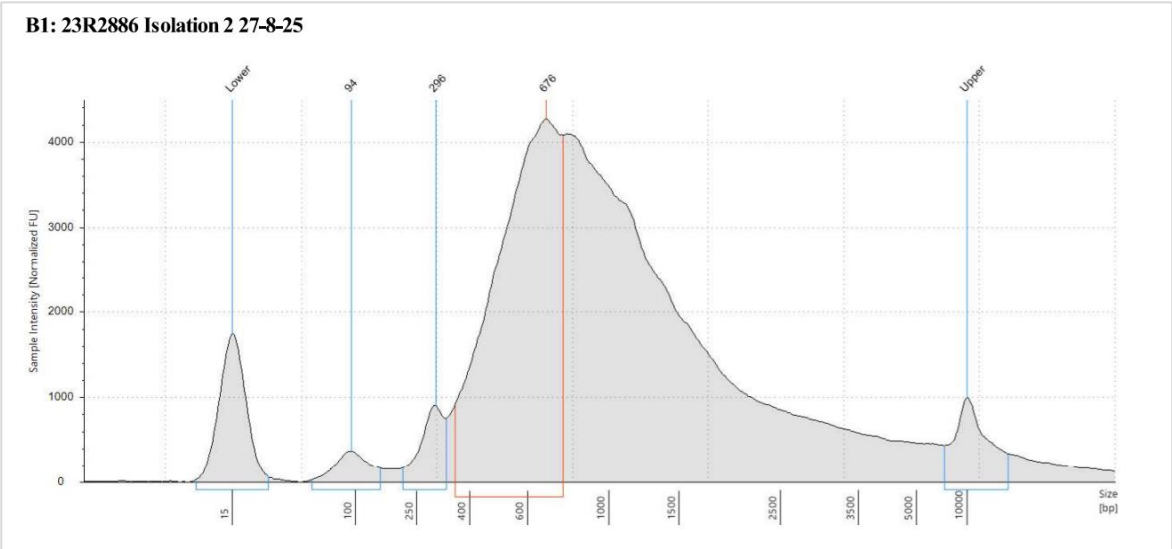


Figure 4.6 Tapestation electropherogram post cDNA amplification step of GEM-X 5' protocol (step 2.4) for sample 1. *A broad peak is seen between 400 – 2000bp.*

Sample 2 (23R2886b)

The concentration of nuclei was 7.89 x10^6 cells/ and the cell viability was 0. For this sample, 6,000 nuclei were loaded onto the GEM-X chip with a capture of approx. 80% of nuclei resulting in aimed recovery of 4800 nuclei. Tapestation was used to measure fragment size and cDNA concentration following cDNA amplification. As with the previous sample a peak was seen on the tapestation electropherogram (Figure 4.7) between 400-2000 bp, with a concentration of 4760 pg/µl. The overall concentration of the sample was 2790 pg/µl, 116ng. 10ng of cDNA was input for ONT library preparation. Table 4.9 shows results of QC steps through ONT library preparation. A final yield of 143 ng was quantified at the end of library preparation which allowed for the recommendation of 100 fmol (1kb, 62 ng) of library to be sequenced.

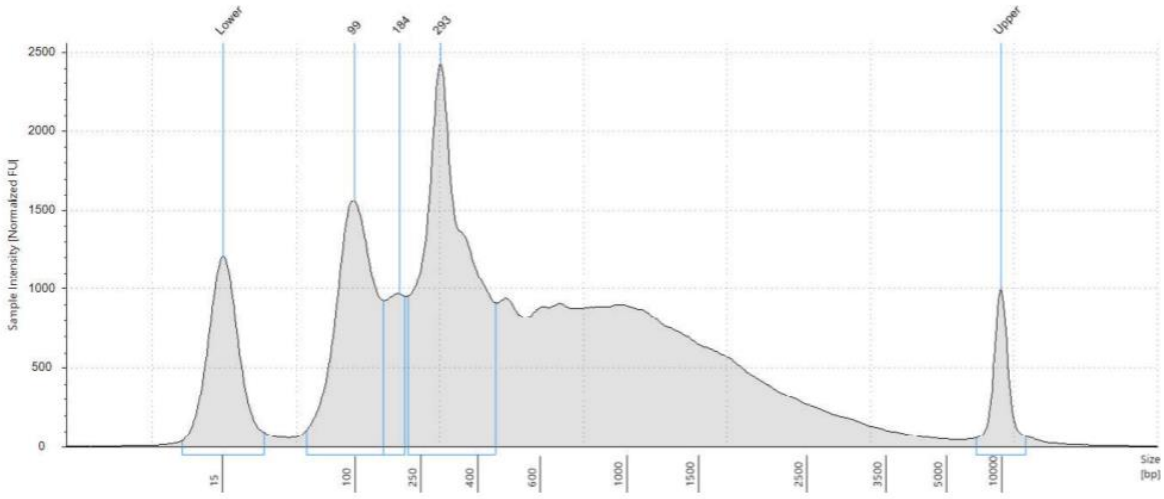


Figure 4.7 Tapestation electropherogram post cDNA amplification step of GEM-X 5' protocol (step 2.4) for sample 2. *A broad peak is seen between 400 – 2000bp.*

Sample 3 (010156)

The concentration of nuclei was 8.67 x10^6 cells/ and the cell viability was 0.1. For this sample, 6,000 nuclei were loaded onto the GEM-X chip. Tapestation was used to measure fragment size and cDNA concentration following cDNA amplification. As with the previous samples a peak was seen on the tapestation electropherogram (Figure 4.8) between 400-2000 bp, with a concentration of 322 pg/µl. The overall concentration of the sample was 322 pg/µl, 13ng. 10ng of cDNA was input for ONT library preparation. Table 4.9 shows results of QC steps through ONT library preparation. A final yield of 128 ng was quantified at the end of library preparation which allowed for the recommendation of 100 fmol (1kb, 62 ng) of library to be sequenced.

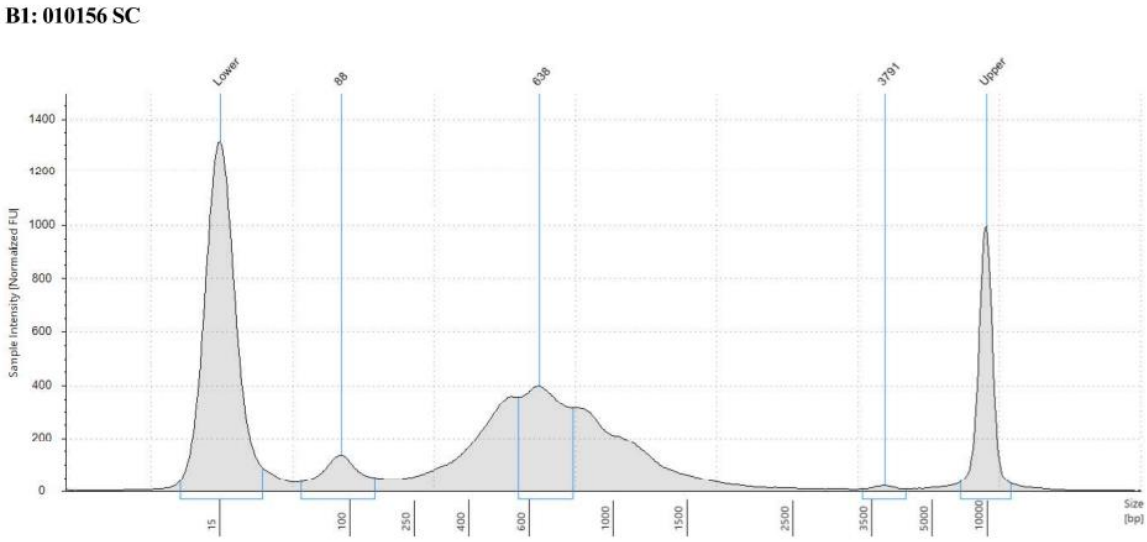


Figure 4.8 Tapestation electropherogram post cDNA amplification step of GEM-X 5' protocol (step 2.4) for sample 3. *A broad peak is seen between 400 – 2000bp.*

4.4.7 Long-read single nuclei 23R2886 shallow sequencing (Sample 1)

Before applying quality control metrics, the data set contained 33,707 cells with a mean of 2,274 reads per cell. The median number of genes per cell (nFeature) was 447 and median UMI's per cell (nCount) 563.

Following applying quality control thresholds, cells with an nFeature of less than 200 or greater than 2500 were excluded from analysis as cells with less than 200 genes per cell are likely low quality and cells with greater than 5,000 genes per cell likely represent doublets. The mean number of genes per cell from the quality controlled data set was 905 (range 342 – 2495) and the mean UMI's per cell was 878 (range 500 – 1585). The percentage of gene counts in a single cell that map to mitochondrial genes is 0%.

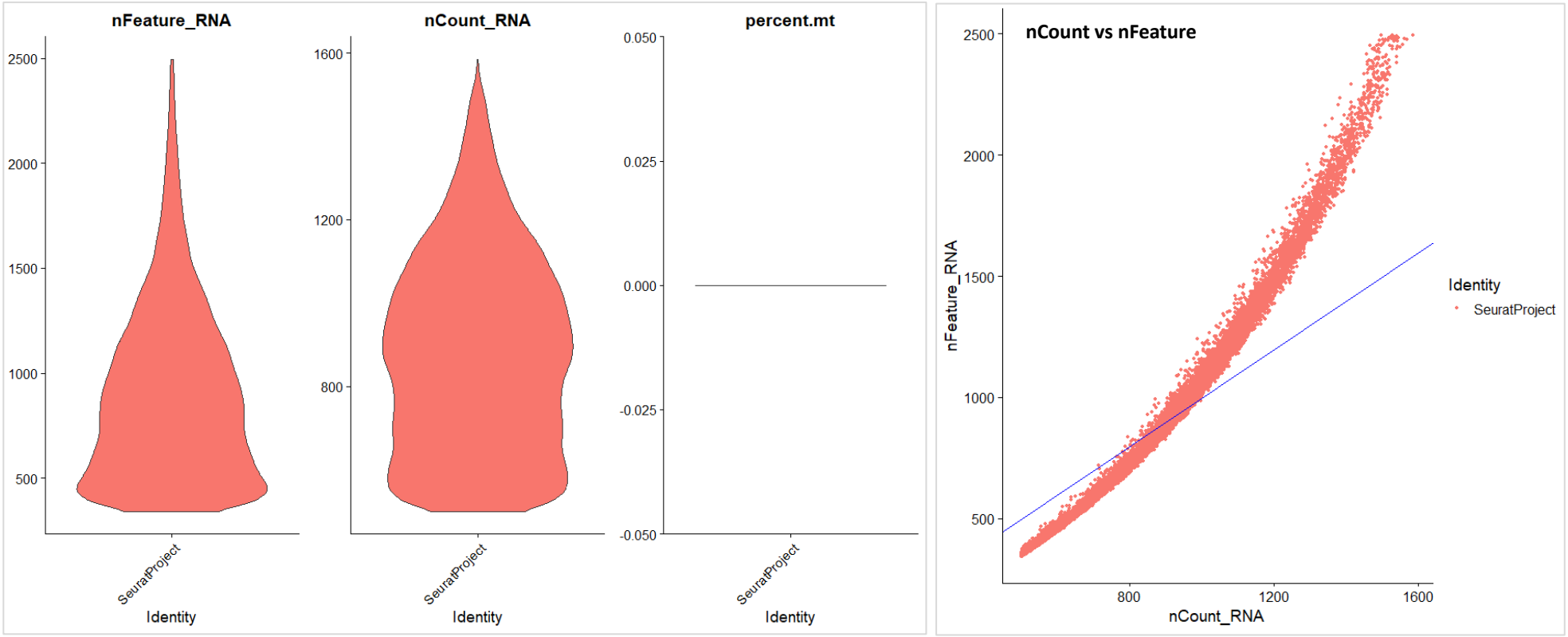


Figure 4.9 Quality control metrics for single cell data for sample 23R2866 shallow read depth. Left: *Violin plots showing the distribution of the number of detected genes per cell (nFeature_RNA), total UMI counts per cell (nCount_RNA), and the percentage of mitochondrial transcripts (percent.mt). across all cells in the Seurat object.* ***Right:*** *Scatter plot of nFeature_RNA versus nCount_RNA with a linear regression line illustrating the positive relationship between sequencing depth and gene detection per cell. These metrics were used to assess the data quality and guide cell filtering.*

Figure 4.9 violin plot shows the distribution of nFeature, nCount and percentage of mitochondrial genes and nFeature vs nCount correlation graph.

An elbow plot (Figure A4 appendix 9) was used to identify the number of principle components; the plot shows the standard deviation of each principal component and the point where the elbow forms (bend in the graph) is identified as the threshold for variation. For this case the number of principle components was identified as 12. The resolution for the UMAP (Figure 4.10) for this case is 0.2 as this represent the general population of the cells. UMAPs at resolution 0.1 – 1 are in Figure A5 in appendix 10. The UMAP at 0.2 resolution shows 6 cell clusters. The large central cluster is formed of 2 smaller clusters, 0 and 1. Clusters 3 represents a smaller and more isolated groups. Cluster 2, 4 and 5 form a smaller group of cells that are close to the large cluster made up of 0 and 2.

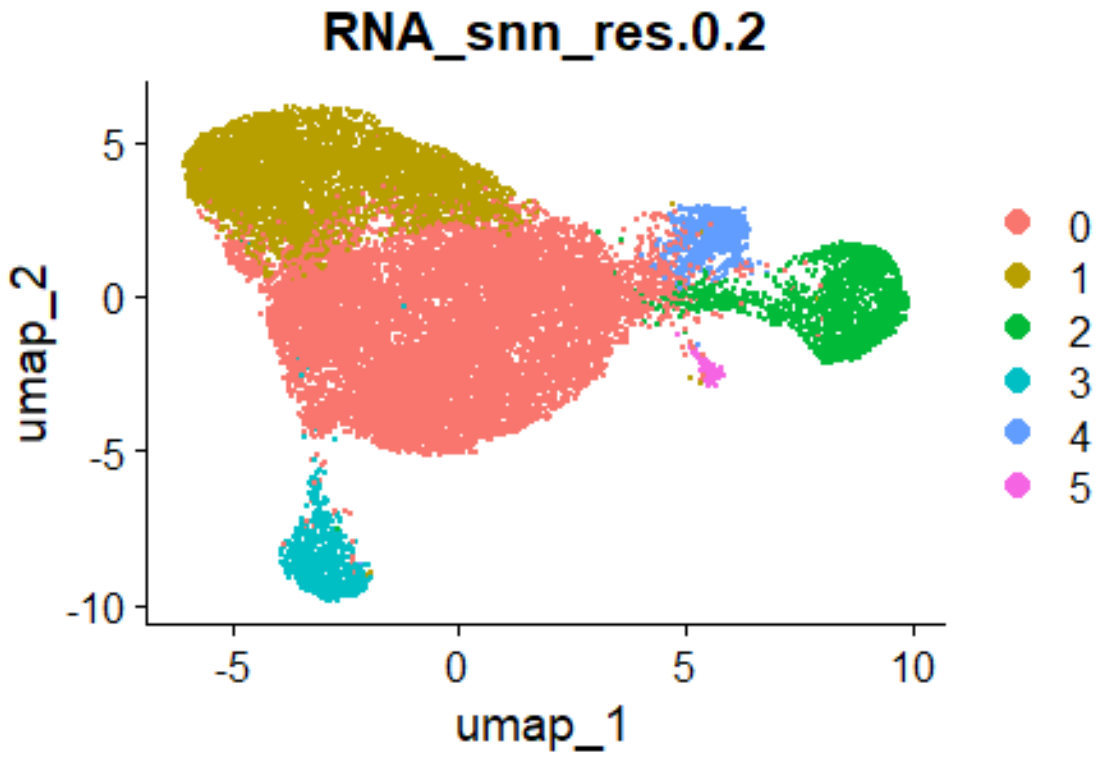


Figure 4.10 UMAP visualisation of single-cell RNA-seq data clustered at resolution 0.2*. Each point represents an individual cell, positioned by UMAP dimensions 1 and 2 based on transcriptomic similarity. Cells are clustered which demonstrates relationships among major cell populations.*

Cells were annotated from the single cell data set compromising of 19,021 after quality control. Cell identities were assigned by applying label transfer from the two reference data sets: Human Limb Embryo Atlas (atlas1) and Endochondral Ossification atlas (2). Using atlas 1, the analysis identified 12 cell types, with Mesenchymal cells (82.28%), fibroblast of dermis (12.78%) and chondrocytes (1.97%) being the most abundant populations. Using atlas 2, 26 cell types were identified with the most prominent cells types being T-cells (48.1%), proximal mesenchyme 16.72%, chondroprogenitor (14.52%), PAX7 myoprogenitor (5.25%), HEY1 osteoblasts (3.34%), smooth muscle cell (1.31%), pre-osteoblasts (1.19%), and prehypertrophic chondrocytes (1.14%).

A cell annotation score between 0-1 with 1 being the highest is applied to each cell. The annotation scores from both atlas 1 and atlas 2 was compared to generate a best prediction cell annotation. This identified 29 cell types including Mesenchymal cell (50.82%), T-cells (22.56%), vascular endothelium (5.65%), Chondroprogenitor (5.50%), Fibroblast of dermis (2.73%), PAX7 myoprogenitor (1.68%), and HEY1 osteoblast (1.42%) being the most prominent cell populations. Annotated cell types are shown in a UMAP (Figure 4.11).

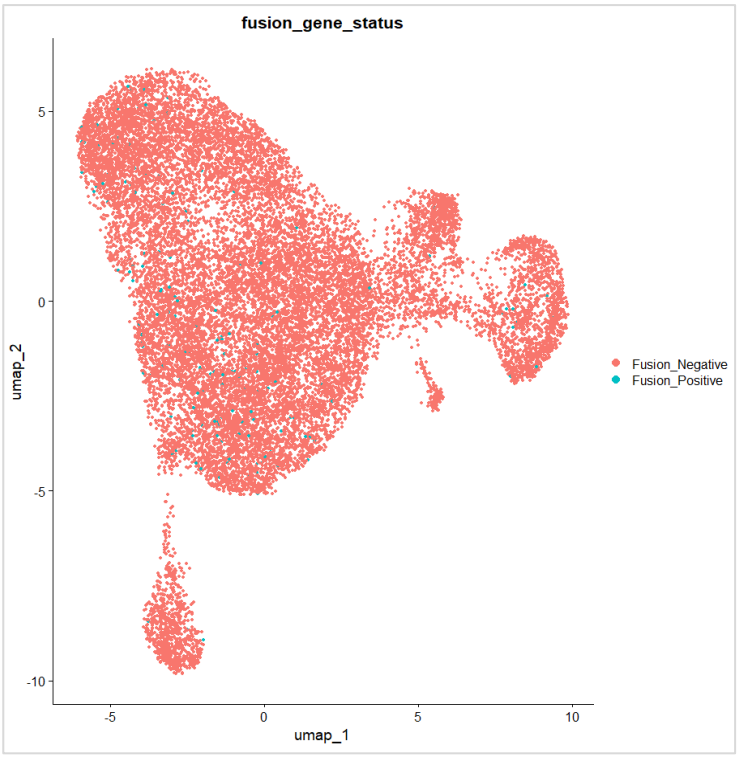
The UMAP for atlas 1 shows one large cluster and three smaller cluster all demonstrating heterogenicity but is dominated by mesenchymal cells. Atlas 2 shows a large cluster mainly dominated by immune cell populations with and 2 smaller cluster: one cluster of smooth muscle cells and one of vascular endothelium cells is more distinct from the main cluster. The best prediction UMAP shows a large cluster which again is heterogenous but has mesenchymal and T-cell populations. Again, there is a cluster of vascular endothelium. The final UMAP shows cells with a confidence score of >0.8. This shows a cluster of endothelial cells and a sparse cluster of T-cells.

Figure 4.11 Comparison of cell type predictions across two reference atlases for 23R2886 shallow read depth (Sample1). *UMAP projections of single-cell transcriptomic data coloured by predicted cell identity.* ***Top left:*** *Predictions obtain using Atlas 1 as a reference showing mesenchymal and connective tissue annotations.* ***Top right:***

Predictions obtained using Atlas 2, showing immune, mesenchymal and endothelial subtypes. ***Bottom left:*** *Integrated best prediction, representing the selected consensus label per cell after comparison across both atlases.* ***Bottom right:*** *subset of cells highlighting confident assigned cell populations (>0.8). Each point represents a single cell.*

After quality control thresholds were applied, *HEY1::NCOA2* fusion was detected in 189 cells. As shown in Figure 4.12 the fusion positive cells are scattered across the UMAP and does not form a distinct cluster. When applying cell annotation, the fusion was detected in a number of cell types.

Using cell annotation from atlas 1, the cells that were fusion positive for *HEY::NCOA2* included mesenchymal cell (80.42%), fibroblast of dermis (15.34%), chondrocytes (1.59%), fibroblast (0.53%), muscle (0.53%), muscle fibroblast (0.53%) osteoblast (0.53%), and skeletal muscle myoblast (0.53%). Using cell annotation from atlas 2, T-cell (31.75%), chondroprogenitor (22.75%), proximal mesenchyme (21.7%), PAX7 myoprogenitor (7.4%), vascular endothelium (4.23%), HEY1 osteoblasts (3.7%), pre-osteoblast (2.12%), pre-hypertrophic chondrocytes (2.12%), erythrocytes (1.59%) and smooth muscle cell (1.59%) cell types were fusion positive. From the best predication cell annotation, the top populations with fusion positive cells include mesenchymal cell (48.15%), T cells (15.34%), chondroprogenitor (11.64%), proximal mesenchyme (8.47%), PAX7 myoprogenitor (3.70%), vascular endothelium (3.70%) fibroblast of the dermis (2.65%), HEY 1 osteoblast (2.12%) and smooth muscle cells (3.70%).



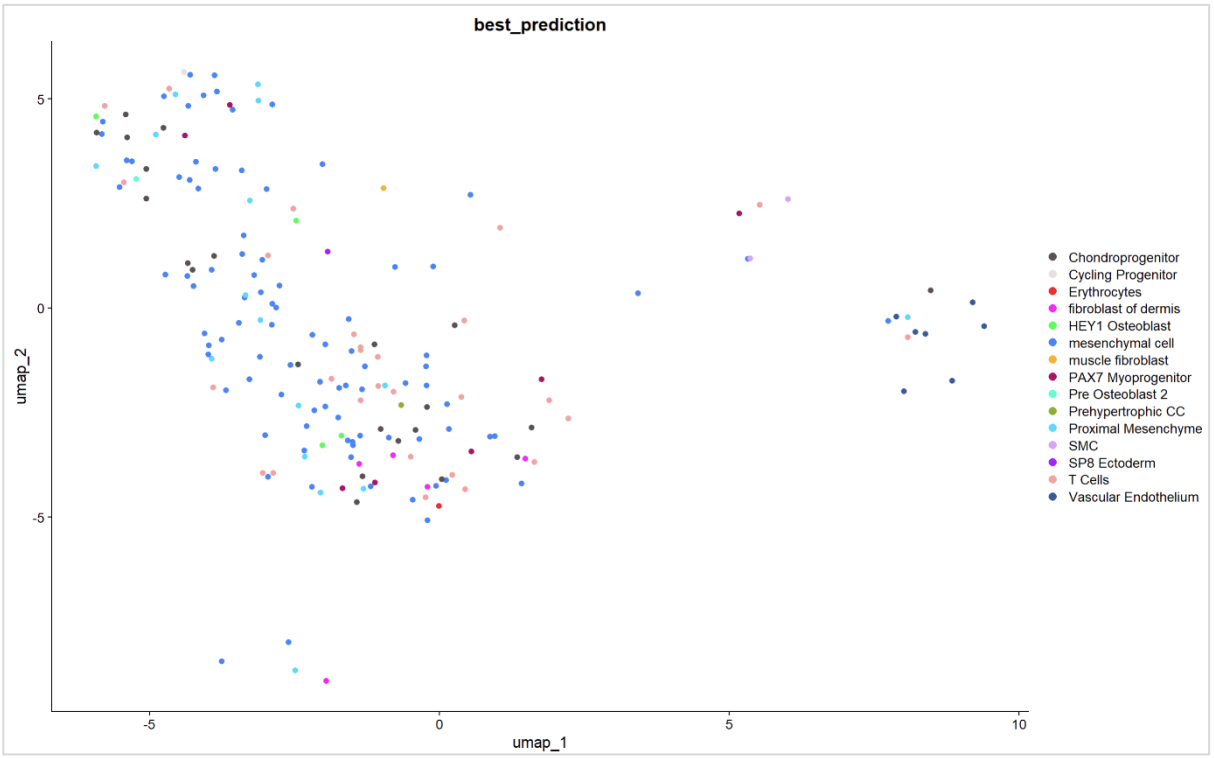


Figure 4.12 UMAP visualisation of fusion gene status for sample 1 23R2886 shallow sequencing. Top: *Each point represents a single cell, coloured by fusion gene stations; fusion negative (red and fusion positive (green). The distribution shows that fusion positive cells are interspersed within the overall cellular landscape without forming a distinct cluster.* ***Bottom:*** *UMAP displaying the best predicted cell annotation. Each colour represents a distinct predicted cell population.*

4.4.8 Long-read single nuclei 23R2886 high read depth (Sample 2)

Before applying quality control metrics, the data set contained 6,340 cells with a mean of 23,751 reads per cell. The median number of genes per cell (nFeature) was 1,078 and median UMI's per cell (nCount) 1,716.

Following applying quality control thresholds, cells with an nFeature of less than 200 or greater than 5000 were excluded from analysis as cells with less than 200 genes per cell are likely low quality and cells with greater than 5,000 genes per cell likely represent doublets. The mean number of genes per cell from the quality controlled data set was 1812 (range 353 – 4998) and the mean UMI's per cell was 1194 (range 501 – 1940). The percentage of gene counts in a single cell that map to mitochondrial genes is 0%.

Figure 4.13 violin plot shows the distribution of nFeature, nCount and percentage of mitochondrial genes and nFeature vs nCount correlation graph.

An elbow plot (Figure A4 appendix 9) was used to identify the number of principle components. For this case the number of principle components was identified as 11. The resolution for the UMAP (Figure 4.14) for this case is 0.2 as this represent the general population of the cells. UMAPs at resolution 0.1 – 1 are in Figure A5 of appendix 10. The UMAP at 0.2 resolution shows 6 cell clusters. The large central cluster is formed of 3 smaller clusters, 0, 1 and 2. Clusters 3, 4 and 5 represent smaller and more isolated groups.

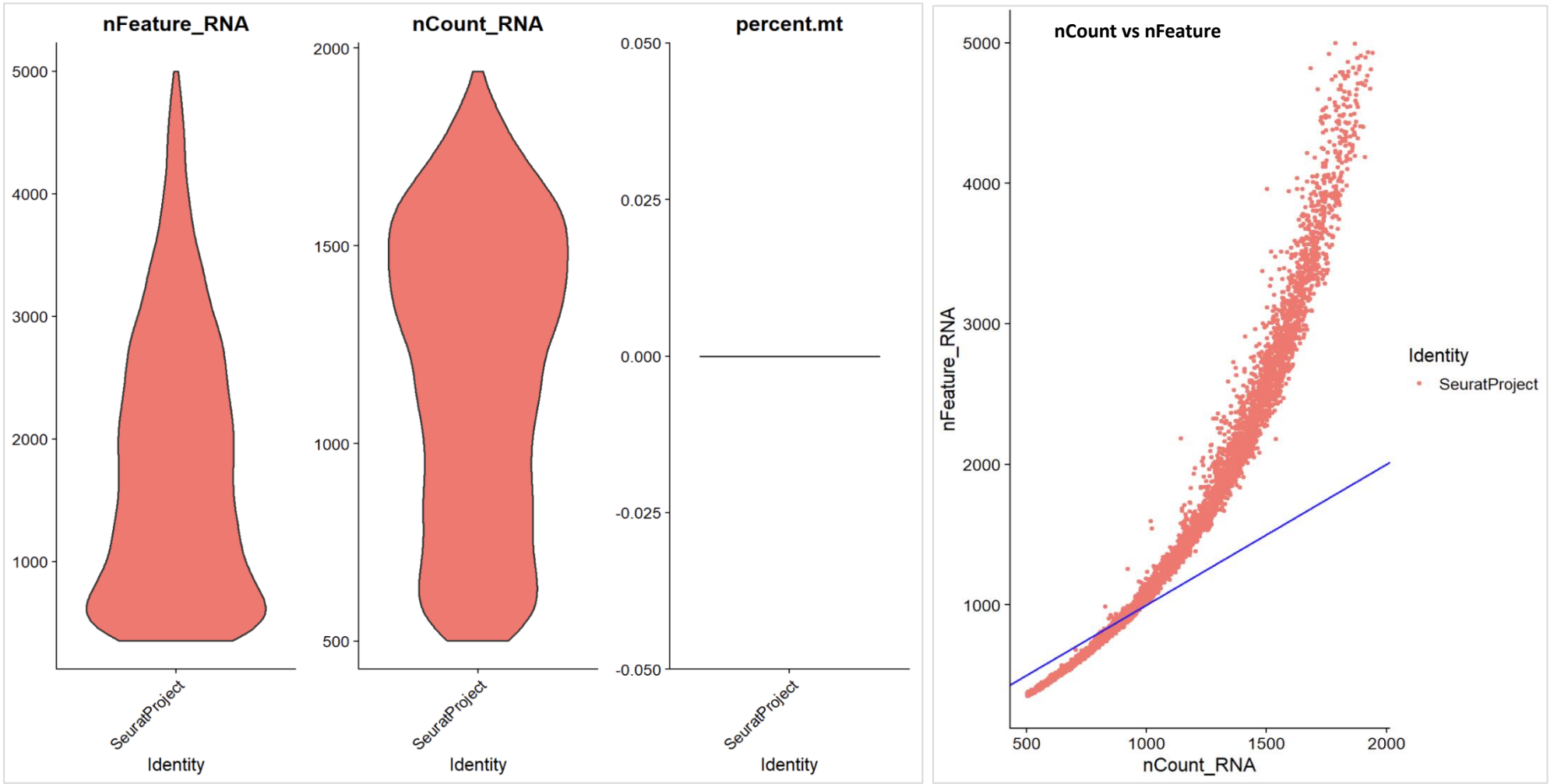


Figure 4.13 Quality control metrics for single cell data for sample 23R2866 high read depth. Left: *Violin plots showing the distribution of the number of detected genes per cell (nFeature_RNA), total UMI counts per cell (nCount_RNA), and the percentage of mitochondrial transcripts (percent.mt). across all cells in the Seurat object.* ***Right:*** *Scatter plot of nFeature_RNA versus nCount_RNA with a linear regression line illustrating the positive relationship between sequencing depth and gene detection per cell. These metrics were used to assess the data quality and guide cell filtering.*

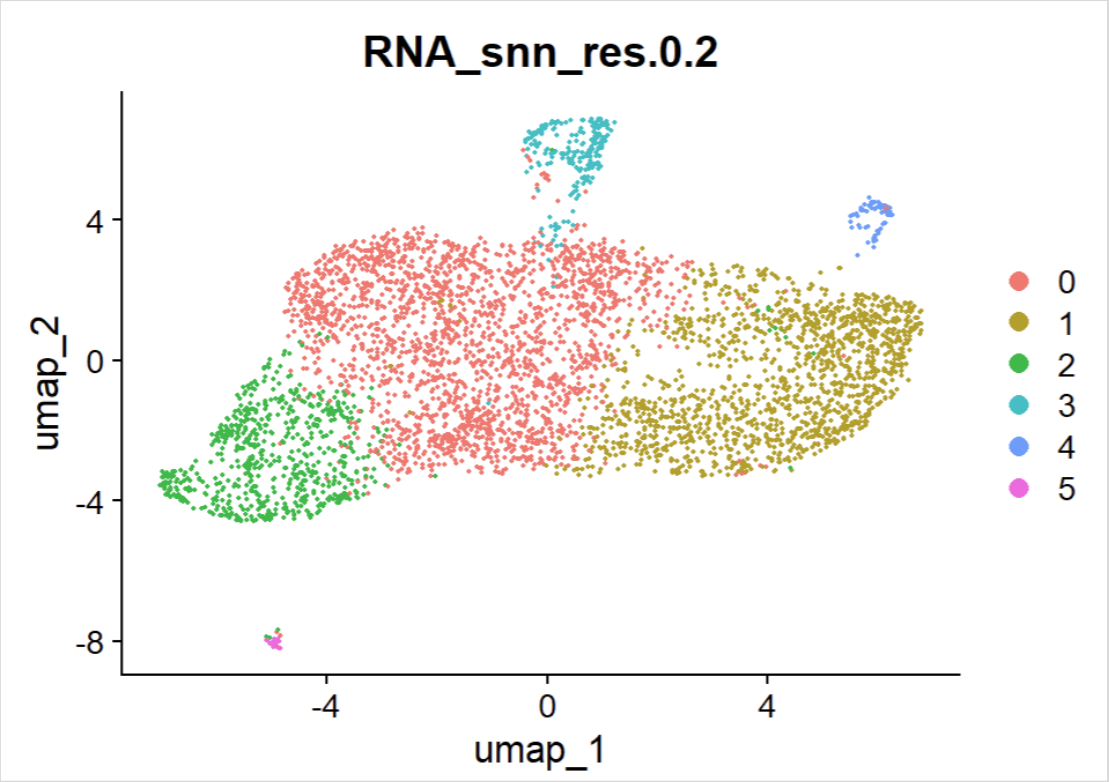


Figure 4.14 UMAP visualisation of single-cell RNA-seq data clustered at resolution 0.2*. Each point represents an individual cell, positioned by UMAP dimensions 1 and 2 based on transcriptomic similarity. Cells are clustered which demonstrates relationships among major cell populations.*

Cells were annotated from the single cell data set compromising of 4,518 after quality control. As per the previous sample, cell identities were assigned by applying label transfer from the two reference data sets: Human Limb Embryo Atlas (atlas1) and Endochondral Ossification atlas (2). Using atlas 1, the analysis identified 16 cell types, with mesenchymal cell (65.18%), fibroblast of dermis (11.47%), chondrocyte (4.98%), primitive red blood cell (3.72%), B-cell (3.7%), skeletal muscle myoblast (3.21%), vein endothelial cell (2.04%), muscle fibroblast (1.15%), chondroblast (1.37%), monocyte (1.11%), and muscle cell (1.02%) being the most abundant populations.

Using atlas 2, 25 cell types were identified with the most prominent cells types being t-cells (52.52%), proximal mesenchyme (14.01%), chondroprogenitor (9.65%), vascular endothelium (8.47%), HEY1 osteoblast (6.68%), PAX7 myoprogenitor (1.55%), smooth

muscle cells (1.48%), prehypertrophic chondrocytes (1.37%), pre-osteoblasts (1.24%), and Sinusoidal Endothelium (1.21%).

A best predication annotation was generated from the highest annotation score for individual cells. This identified 29 cell types including Mesenchymal cell (37.25%), T-cells (32.09%), vascular endothelium (6.71%), proximal mesenchyme (5.82%), Chondroprogenitor (3.83%), HEY1 osteoblast (3.12%), fibroblast of dermis (2.05%), skeletal muscle myoblast (1.90%) and sinusoidal endothelium (1.08%). Annotated cell types are shown in a UMAP (Figure 4.15).

The UMAP for atlas 1 shows one large cluster and two smaller cluster all demonstrating heterogenicity but is dominated by mesenchymal cells. Atlas 2 is similar to atlas 1 with a large cluster mainly dominated by T-cells with and 2 smaller cluster: one cluster of smooth muscle cells and one of vascular endothelium cells is more distinct from the main cluster. The best prediction UMAP shows a large cluster which again is heterogenous but has mesenchymal and T-cell populations. Again, there is a cluster of vascular endothelium. The final UMAP shows cells with a confidence score of >0.8. This shows three clusters of cells, with large cluster dominated by T-cells, and a smaller cluster of smooth muscle cells and endothelial cells.

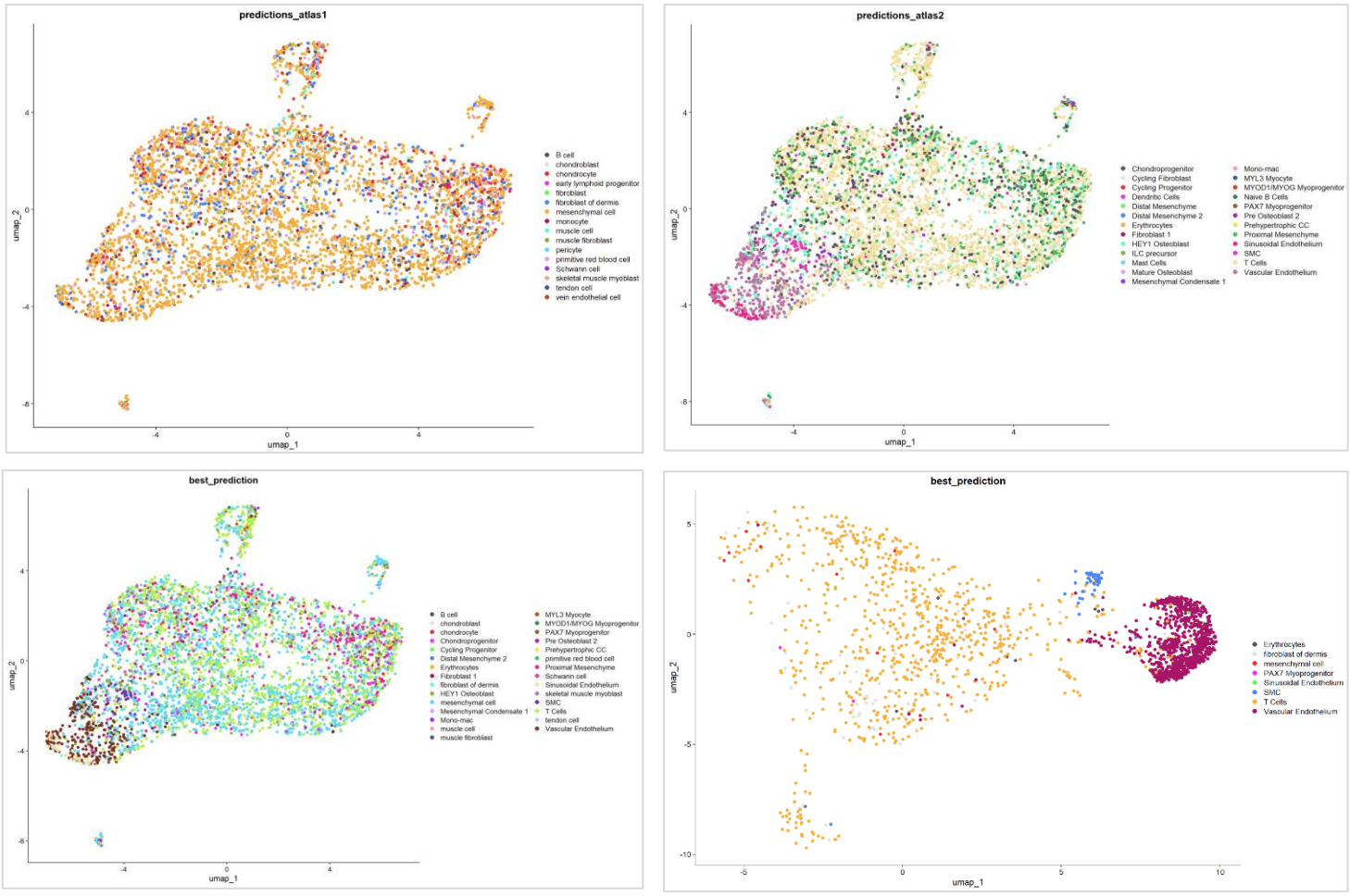


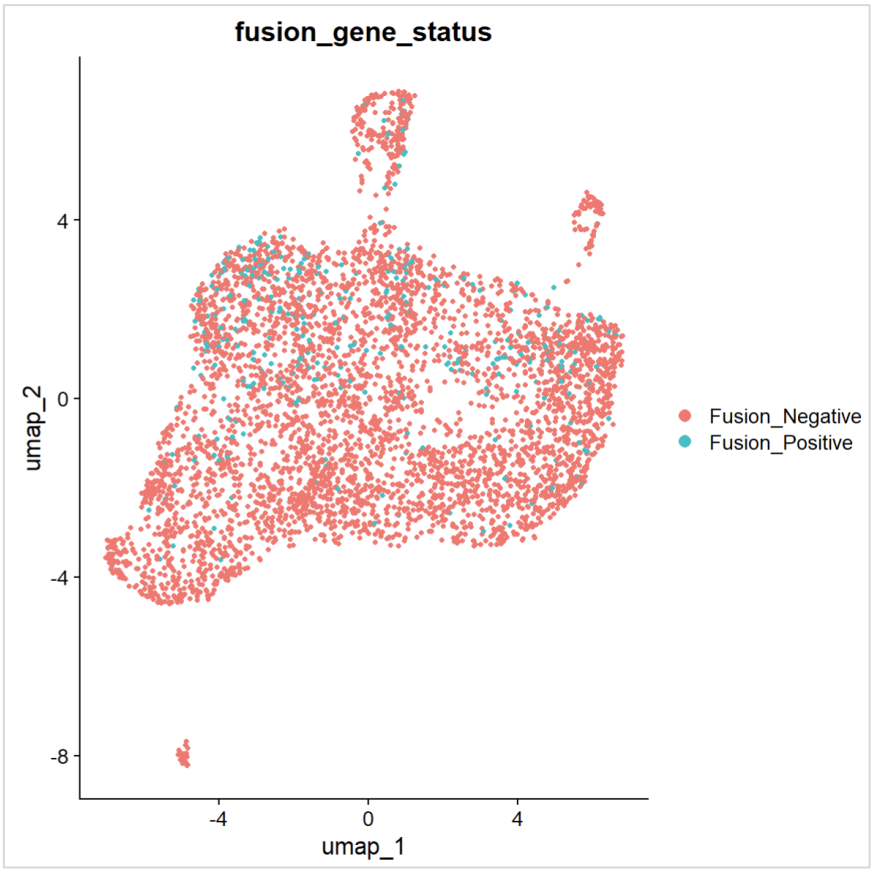
Figure 4.15 UMAP projections of single-cell transcriptomic data coloured by predicted cell identity for 23R2886 high read depth. *Top left: Predictions obtain using Atlas 1 as a reference showing mesenchymal and connective tissue annotations. Top right: Predictions obtained using Atlas 2, showing immune, mesenchymal and endothelial subtypes. Bottom left: Integrated best prediction, representing the selected consensus label per cell after comparison across both atlases. Bottom right: subset of cells highlighting confident assigned cell populations (>0.8). Each point represents a single cell.*

Following quality control thresholds being applied, *HEY1::NCOA2* was detected in 328 cells. As shown in Figure 4.16 the fusion positive cells are scattered across the UMAP and does not form a distinct cluster.

When applying cell annotation, as with the shallow sequenced sample, the fusion was detected in several cell types. Using cell annotation from atlas 1, the cells that were fusion positive for *HEY::NCOA2* included mesenchymal cell (53.96%), fibroblast of dermis (14.02%), chondrocytes (7.01%), B-cell (4.27%), skeletal muscle myoblast (6.40%), primitive red blood cells (3.66%), vein endothelium (2.44%), Chondroblast (2.13%), monocyte (1.52%), and Schwann cell (1.52%).

Using cell annotation from atlas 2, T-cell (46.03%), proximal mesenchyme (18.60%), chondroprogenitor (13.11%), HEY1 osteoblasts (7.62%), pre-hypertrophic chondrocytes (4.27%), pre osteoblast (3.05%), vascular endothelium (3.04%), and smooth muscle cell (1.22%).

From the best predication cell annotation, the top populations with fusion positive cells include T cells (30.18%), mesenchymal cell (28.66%), proximal mesenchyme (10.37%), chondroprogenitor (6.10%), skeletal muscle myoblast (4.57%), HEY 1 osteoblast (3.96%), fibroblast of the dermis (3.05%), vascular endothelium (2.44%), prehypertrophic chondrocytes (2.44%), primitive red blood cells (1.83), and Schwann cell (1.52%).



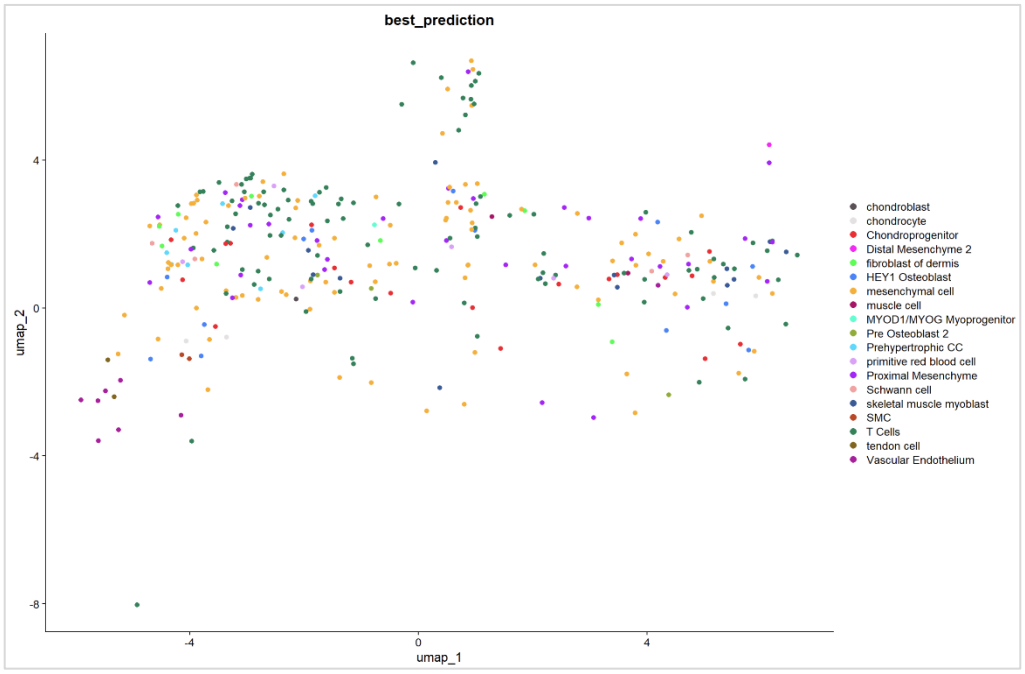


Figure 4.16 UMAP visualization of fusion gene status and predicted cell identities for sample 2 23R2886 high sequencing depth. Top: *UMAP displaying single cells colour by fusion gene status. Fusion- negative cells are showing in red and fusion-positive in green. Fusion-positive cells are distributed across the clusters.* ***Bottom:*** *UMAP displaying the best predicted cell annotation. Each colour represents a distinct predicted cell population.*

4.4.9 Long-read single nuclei 010156 – (Sample 3)

Before applying quality control metrics, the data set contained 8,605 cells with a mean of 23,929 reads per cell. The median number of genes per cell (nFeature) was 759 and median UMI's per cell (nCount) 1,059.

Following applying quality control thresholds, cells with an nFeature of less than 200 or greater than 3500 were excluded from analysis as cells with less than 200 genes per cell are likely low quality and cells with greater than 3,500 genes per cell likely represent doublets. The mean number of genes per cell from the quality controlled data set was 978 (range 344 – 3496) and the mean UMI's per cell was 978 (range 500 – 1772). The percentage of gene counts in a single cell that map to mitochondrial genes is 0%.

Figure 4.17 violin plot shows the distribution of nFeature, nCount and percentage of mitochondrial genes and nFeature vs nCount correlation graph.

As previously stated in section 4.4.6 an elbow plot (Figure A4 of appendix 9) was used to identify the number of principle components. For this case the number of principle components was identified as 9. The resolution for the UMAP (Figure 4.18) for this case is 0.2 as this represent the general population of the cells. UMAPs at resolution 0.1 – 1 are in Figure A5 of appendix 10. The UMAP at 0.2 resolution shows 6 cell clusters. The large central cluster is formed of 4 smaller clusters, 0, 1, 2 and 4. Clusters 3 and 5 represent smaller and more isolated groups.

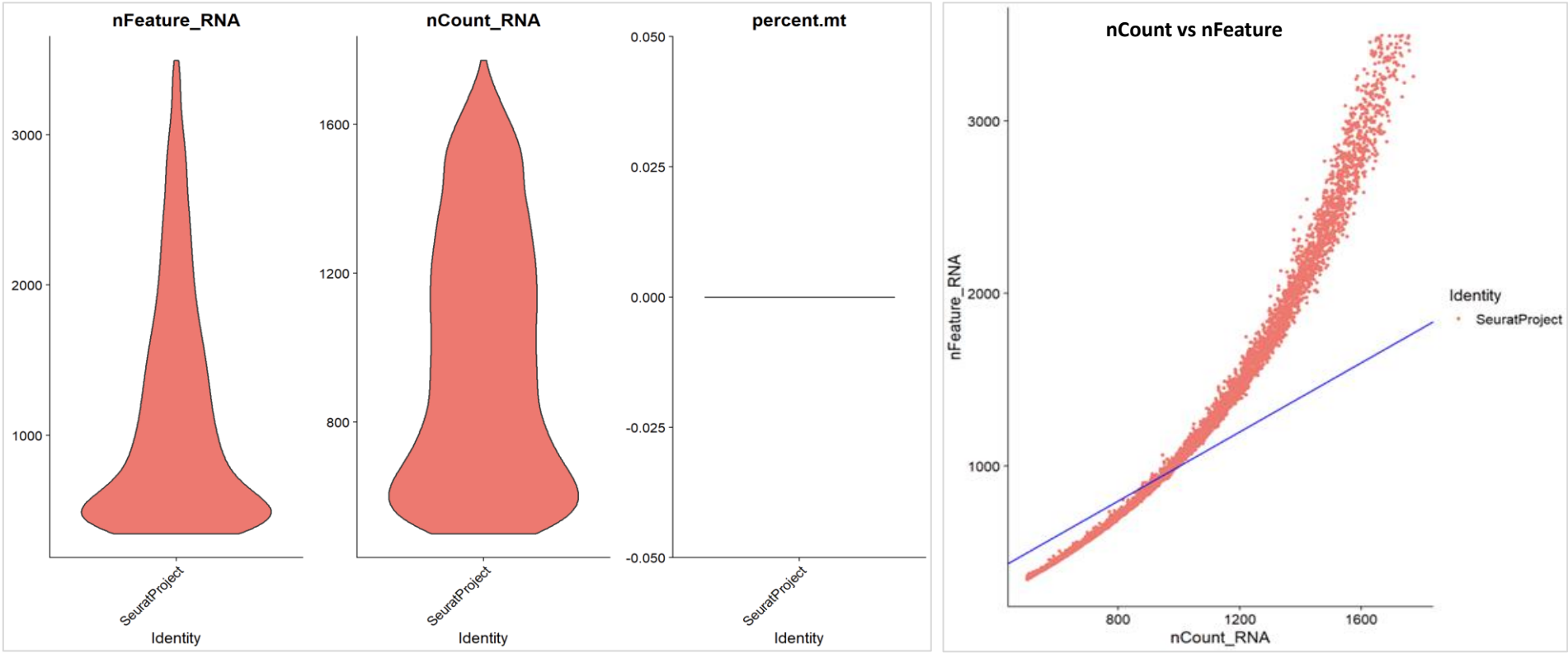


Figure 4.17 Quality control metrics for single cell data for sample 010156 high read depth. Left: *Violin plots showing the distribution of the number of detected genes per cell (nFeature_RNA), total UMI counts per cell (nCount_RNA), and the percentage of mitochondrial transcripts (percent.mt). across all cells in the Seurat object.* ***Right:*** *Scatter plot of nFeature_RNA versus nCount_RNA with a linear regression line illustrating the positive relationship between sequencing depth and gene detection per cell. These metrics were used to assess the data quality and guide cell filtering.*

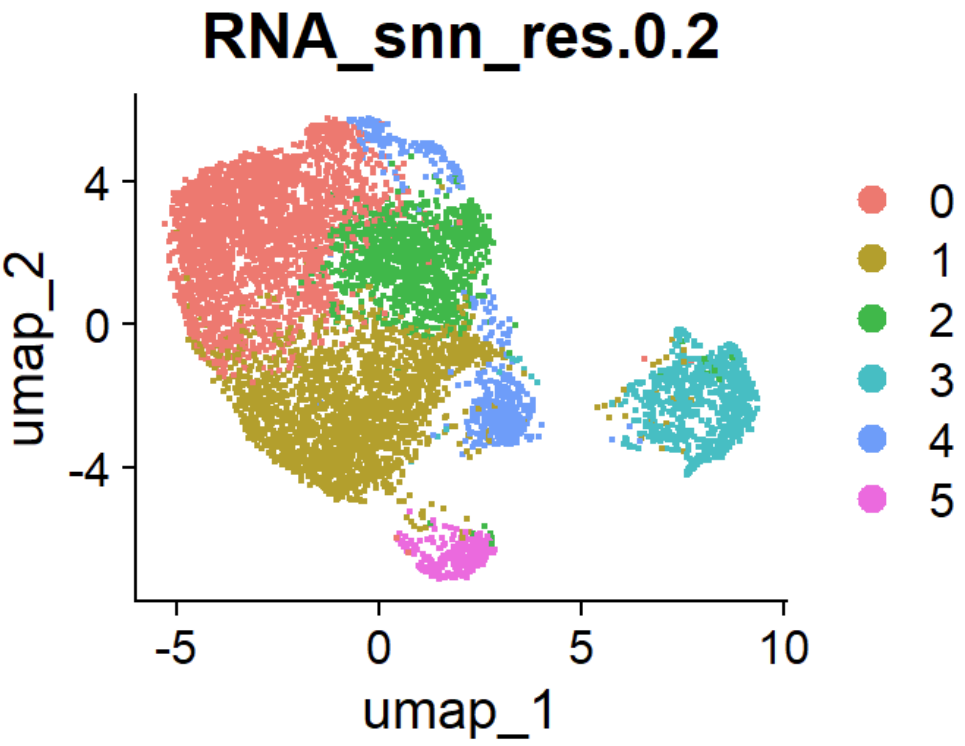


Figure 4.18 UMAP visualisation of single-cell RNA-seq data clustered at resolution 0.2*. Each point represents an individual cell, positioned by UMAP dimensions 1 and 2 based on transcriptomic similarity. Cells are clustered which demonstrates relationships among major cell populations.*

Cells were annotated from the single cell data set compromising of 7,114 after quality control. As per the previous samples, cell identities were assigned by applying label transfer from the two reference data sets: Human Limb Embryo Atlas (atlas1) and Endochondral Ossification atlas (2). Using atlas 1, the analysis identified 15 cell types, with mesenchymal cell (61.1%), muscle fibroblast (12.4%), early lymphoid progenitor (6.47%), chondrocytes (3.07%), mast cell (2.67%), basal cell (2.35%), skeletal muscle myoblast (2%), macrophage (1.45%), and skeletal muscle myoblast (1.34%) being the most abundant populations.

Using atlas 2, 15 cell types were identified with the most prominent cells types being T-cells (43.67%), proximal mesenchyme (23.43%), chondroprogenitor (14.55%), vascular

endothelium (7.87%), HEY1 osteoblast (4.89%), PAX7 myoprogenitor (2.14%), and smooth muscle cells (1.93%).

A best predication annotation was generated from the highest annotation score for individual cells. This identified 25 cell types including Mesenchymal cell (42.73%), T-cells (18.92%), proximal mesenchyme (12.13%), Chondroprogenitor (8.22%), vascular endothelium (6.83%), HEY1 osteoblast (2.64%), mast cell (2.42%), and smooth muscle cell (1.08%). Annotated cell types predictions from atlas 1, atlas 2, and best predication is showing in Figure 4.19. The UMAP for atlas 1 shows one large cluster and three smaller cluster all demonstrating heterogenicity. Atlas 2 shows a large cluster mainly dominated by T-cells with and 2 smaller cluster: one cluster of smooth muscle cells and one of vascular endothelium cells. The best prediction UMAP shows a large cluster which again is heterogenous but has mesenchymal and immune group populations. Again, there are two smaller clusters, one of smooth muscle cells and the other of vascular endothelium. The final UMAP shows cells with a confidence score of >0.8. This shows three clusters of cells, with large cluster dominated by mesenchymal cells, and a smaller cluster of smooth muscle cells and endothelial cells.

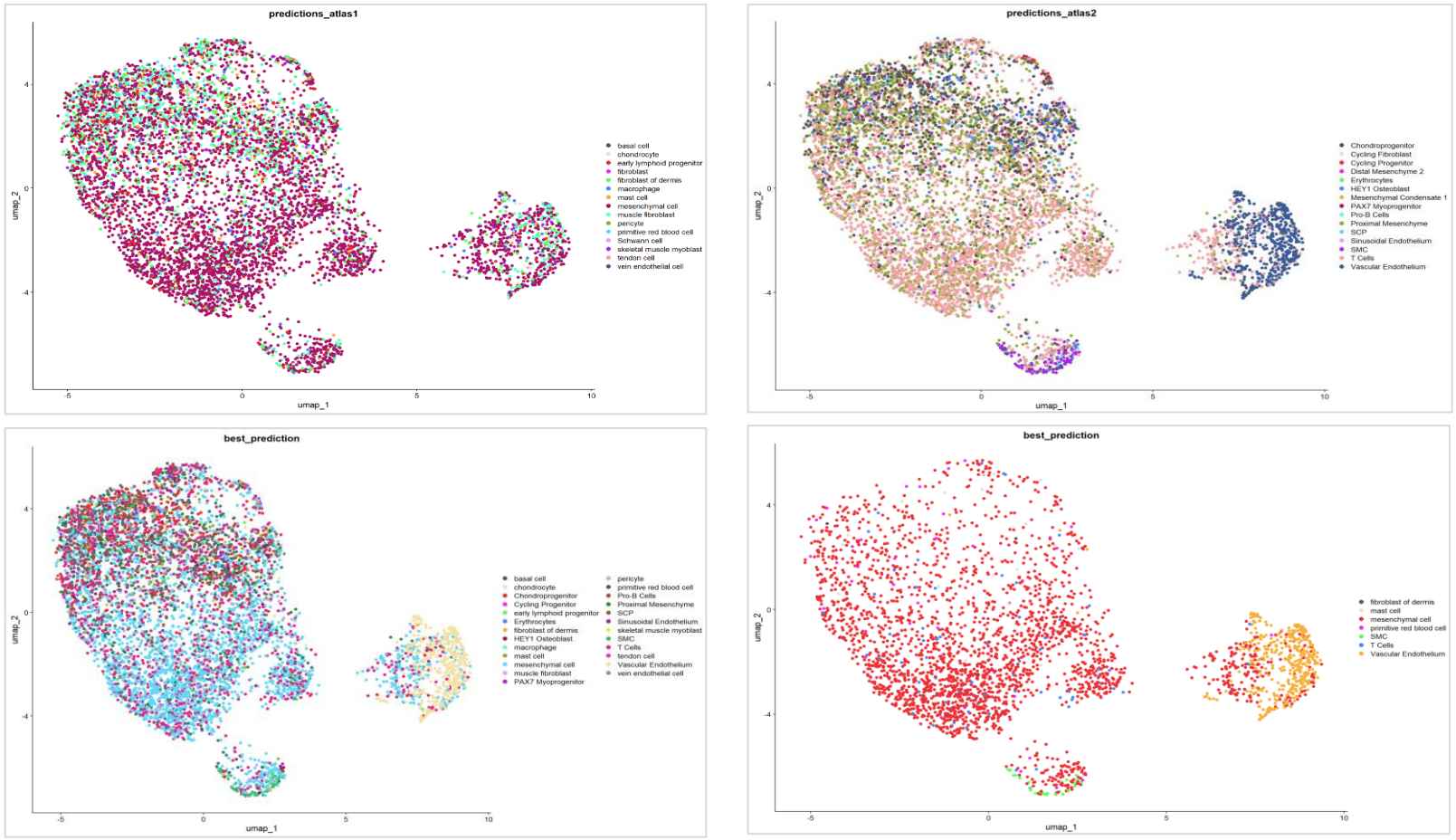


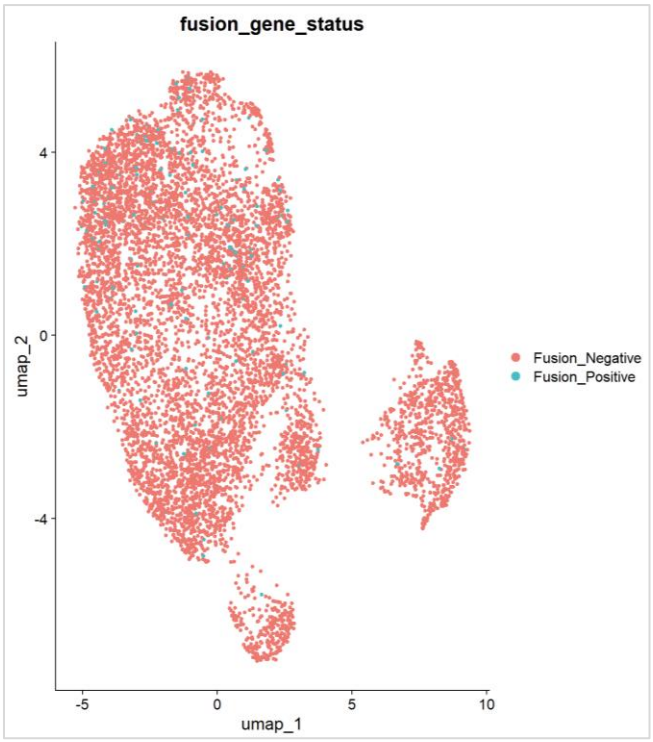
Figure 4.19 UMAP projections of single-cell transcriptomic data coloured by predicted cell identity for 0101056 high read depth. *Top left: Predictions obtain using Atlas 1 as a reference showing mesenchymal annotations. Top right: Predictions obtained using Atlas 2, showing immune, mesenchymal and endothelial subtypes. Bottom left: Integrated best prediction, representing the selected consensus label per cell after comparison across both atlases. Bottom right: subset of cells highlighting confident assigned cell populations (>0.8). Each point represents a single cell.*

Following quality control thresholds being applied, *HEY1::NCOA2* was detected in 151 cells. Figure 4.20 shows the distribution of fusion positive cells vs fusion negative. the fusion positive cells are scattered across the UMAP and does not form a distinct cluster.

When applying cell annotation, as with the previous samples, the fusion was detected in several cell types. Using cell annotation from atlas 1, the cells that were fusion positive for *HEY::NCOA2* included mesenchymal cell (52.32%), muscle fibroblast (16.56%), fibroblast of dermis (9.93%), early lymphoid progenitor (6.47%), chondrocytes (5.96%), mast cell (3.31%), basal cell (1.98%), primitive red blood cell (1.32%) and skeletal muscle myoblast (1.32%).

Using cell annotation from atlas 2, T-cell (31.79%), proximal mesenchyme (23.43%), chondroprogenitor (27.15%), HEY1 osteoblasts (5.3%), PAX7 myoprogenitor (3.31%), and vascular endothelium (1.98%).

From the best predication cell annotation, the top populations with fusion positive cells include mesenchymal cell (35.75%), chondroprogenitor (17.22%), T-cell (16.56%), proximal mesenchyme (14.57%), HEY1 osteoblast (3.31%), mast cell (3.31%), vascular endothelium (1.99%), fibroblast of dermis (1.32%), PAX7 myoprogenitor (1.32%), and primitive red blood cell (1.32%).



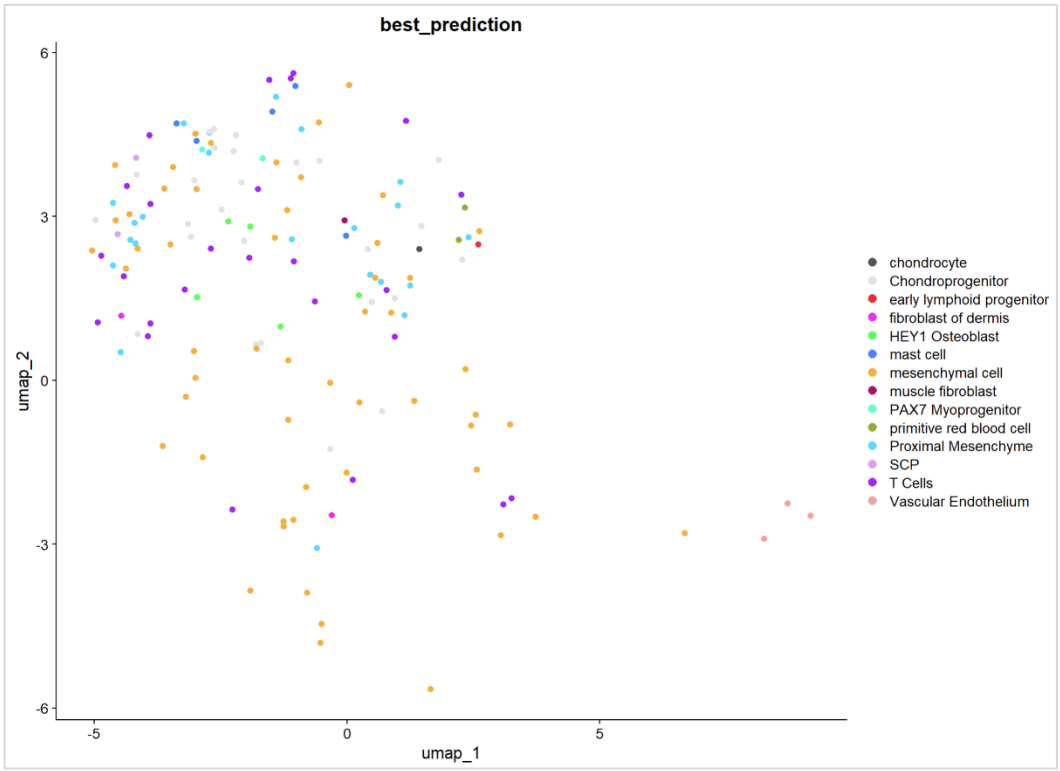


Figure 4.20 UMAP visualization of fusion gene status and predicted cell identities for sample 3 010156 high sequencing depth. Top: *UMAP displaying single cells colour by fusion gene status. Fusion- negative cells are showing in red and fusion-positive in green. Fusion-positive cells are distributed across the clusters.* ***Bottom:*** *UMAP displaying the best predicted cell annotation. Each colour represents a distinct predicted cell population*

4.5 Discussion

Comparison of Long-read and short-read sequencing platform performance and investigation of somatic driver landscape

In this study, the performance of short-read and long-read whole genome sequencing for the detection of the fusion gene was assessed. Overall, there was high concordance between the two sequencing platforms for the detection of the fusion gene and the detection of potentially clinically relevant somatic alterations.

The *HEY1::NCOA2* fusion gene, which characterises MCS, was detected in 88% of cases by both short-read and long-read sequencing, demonstrating that both approaches are effective at identifying clinically relevant fusions. Similarly, the *EWSR1::NFATC2* fusion was detected by both platforms in the single affected case, further supporting the reliability of both platforms in identifying chromosomal rearrangements. The *HEY1::NCOA2* fusion is caused by a 10-megabase deletion which would usually require a sequencing platform that can sequence reads that span the fusion breakpoint which is a known challenge for short-read sequencing as the platforms typically generate reads up to 300 base pairs long. These are often too short to span the fusion breakpoint. However, improvement in bioinformatic pipelines have enhanced the ability to detect fusion genes. In this dataset, ESVEE and LINX bioinformatic pipelines, which are part of the oncoanalyser pipeline, were used to detect the fusion. ESVEE (Hartwig medical foundation, 2026) and LINX (Shale *et* al, 2022) are complimentary tools for detecting large structural variants from short-read WGS data. ESVEE identifies the genomic rearrangement breakpoints using splits reads and local assembly. LINX subsequently interprets these variants by clustering the breakpoints and determines if the rearrangements form fusion genes.

One sample (23R2886) had two breakpoints identified by LR, first at coordinates chr8:79765922 // chr8:70145965 which falls into the expected fusion occurring between exon 13 of *NCOA2* and exon 4 of *HEY1.* The second break point (chr8:79764509// chr8:70146039) occurs within exon 5 of *HEY1*. The detection of two breakpoints in one case is unusual and but has been previously reported in a published case study by Kishikawa *et al*, 2024. In this case study the characteristic fusion between exon 4 of *HEY1* and exon 13 of *NCOA2* was detected but in addition, a previously unreported fusion between exon 4 of *HEY1* and exon 14 of *NCOA2*. Whilst the mechanism of the addition fusion is unknown, it is suggested that it could be caused by alternate splicing. Alternative splicing is a post-transcriptional-process in which a pre-mRNA transcript is spliced in alternative ways to include or exclude specific exons, which results in the production of mRNA isoforms (Zhu et al, 2025). However, in the case discussed in this thesis, the presence of two distinct *HEY1::NCOA2* genomic breakpoints detected by ONT rule out the mechanism of alternative splicing and instead indicate multiple independent fusion events.

Accurate detection of the *HEY1::NCOA2* fusion is critical for precise diagnosis and disease classification. In this study, an *EWSR1::NFATC2* fusion was identified in one case that had been diagnosed histopathological as MCS; however, no prior molecular analysis had been performed to confirm the presence of the *HEY1::NCOA2* fusion. The *EWSR1::NFATC2* fusion defines a distinct tumour entity (Szuhai *et al*, 2000) with morphological overlap with both Ewings sarcoma and myxoid chondrosarcoma but no described overlap with MCS. These tumours are classified as Ewing-like sarcoma and are recognised in the 5th edition of WHO classification of tumours: Soft tissue and bone tumour, as a round cell sarcoma with EWSR1-non-ETS fusions (Le Loarer, Szuhai and Tirode, 2020).

The absence of the *HEY1::NCOA2* fusion and presence of a fusion characteristic of a different tumour entity demonstrates the critical role of molecular analysis in establishing an accurate diagnosis, particularly for tumour entities that have similar histological features. An incorrect diagnosis in this context has significant therapeutic implications. Although chemotherapy regimens for MCS are similar to those used for Ewing sarcoma (Remiszewski *et al*, 2025), tumours involving *EWSR1::NFATC2* rearrangements do not respond to standard Ewing sarcoma therapy, (Diaz-Perez *et al*, 2019) and therefore the patient could be directed to incorrect chemotherapy, and risks exposure and side effects to unnecessary cytotoxic treatment. Consequently, failure to perform appropriate molecular testing may result in receiving ineffective treatment.

A further aim of this study was to investigate the somatic driver landscape beyond the fusion gene. Deletions in *ERBB4* were the most frequent genomic alteration identified, detected in 66% by short-read sequencing. *ERBB4* is a receptor kinase that mediates cell proliferation and differentiation. In other cancer types, overexpression of *ERBB4* can act as tumour suppressor which improves patient prognosis but, in some cancers, it can act as an oncogene and promote cell survival (El-Gamal, 2021).

The breakpoint mutations detected by short-read may have been missed by long-read due to the difference in read depth. The average coverage from the short-read sequencing data was 41X and from long-read sequencing data was 18X. In this data set, it may have been that the mutation occurred in a low number of reads and therefore, filtered during bioinformatic analysis.

Other mutations of potential clinical significance identified in this cohort involved the *CHEK2, LEF1, CCDC6,* and *CHD4* genes. A missense variant in *CHEK2* (p.Thr45Met) gene was

detected in one sample using both long-read and short-read sequencing. *CHEK2* is a tumour suppressor gene with a central role in the DNA damage response pathway (Antoni *et al*, 2007). Upon activation by DNA double-strand breaks, *CHEK2* phosphorylates downstream targets, leading to cell-cycle arrest, DNA repair or apoptosis. Loss of CHEK2 function contributes to genomic instability and tumorigenesis, and pathogenic germline variants have been associated with pathogenic germline variants being associated with an increase susceptibility to several cancers, most notably breast cancer (Cybulski *et al*, 2004; Antoni *et al*, 2007). However, mesenchymal chondrosarcoma is not currently recognised as a tumour subtype associated with *CHEK2* mutations. To further assess the potential relevance of these mutations in MCS, whole genome sequencing of matched germline samples would be required to determine if this tumour occurred with a germline *CHEK2* mutation syndrome.

Variants identified in *LEF1* and *CCDC6* may also be of biological relevance in mesenchymal chondrosarcoma. *LEF1* encodes a secreted protein that acts as a regulatory molecule that reduces the activity of the TGF-β signalling pathway which influences the Wnt signalling pathway which is involved in cellular proliferation, differentiation, and cellular apoptosis (Santiago *et al*, 2017). Alterations in *LEF1* have been reported in several cancer types, including lung and prostate cancer, where dysregulation of the signalling pathways contributes to tumour progression (Santiago *et al,* 2017). Although *LEF1* has not been described as a recurrently altered gene in mesenchymal chondrosarcoma, dysregulation of Wnt signalling pathway is a recognised feature of the *HEY1::NCOA2* positive tumours, suggesting that *LEFT1* variants may influence tumour biology.

CCDC6 is a protein-coding gene that plays a key role in DNA damage response, transcriptional regulation. (Morra *et al*, 2017; Morra *et al*, 2021). It is also a known fusion partner in multiple tumour types, where the fusions create oncogenic drivers promoting tumorigenesis (Cerrato *et al*, 2017). For instance, in the *CCDC6::RET* fusion, *CCDC6* promotes RET kinase activate which leads to downstream signalling of MAPK, PI3K/AKT pathways which in turn drive tumour proliferation (Cerrato *et al*, 2017). These findings highlight *CCDC6* fusions as potential therapeutic targets by blocking signalling and inhibiting tumour growth (Laxmi, Gupta and Gupta, 2019). Although *CCDC6* rearrangements were not identified in this cohort and variants in this gene have not been previously reported in MCS, dysregulation of PIK3-AKT pathways is a recognised feature of *HEY1::NCOA2* fusion. Therefore, variants in *CCDC6* could potentially contribute to dysregulated signalling, suggesting that *CCDC6* may represent a potential therapeutic target.

In summary, while these variants in *CHEK2, LEF1,* and *CCDC6* are not yet established drivers in mesenchymal chondrosarcoma, their known roles in DNA damage response and signalling pathways suggest they could contribute to tumour biology. Further functional studies and comprehensive germline analyses will be essential to clarify their clinical significance and potential as therapeutic targets in MCS.

Long-read epigenetic modification

Genome in a Bottle reference sample are well characterized human genomes that are widely used as benchmarking standards (Zook *et al*, 2016). ONT have provided sequencing data for the GIAB for benchmarking ONT technology. While these data sets provide standardised benchmarks for assessing sequencing performance, they are limited in ability to represent patient-specific epigenetic variation. They are suited for the initial comparison

to verify analytical performance, however, for biological relevant methylation analysis, matched germline samples provide a more suitable reference. The ONT long-read data and analysis pipeline was successful in detecting 5mC and 5hmC base modifications. However, this dataset is limited by the absence of patient-matched controls.

In existing literature, the *HEY1* promotor region is not well defined, therefore, in this dataset, up to 3 Kb upstream of *HEY1* was analysed for base modifications following approaches commonly used to capture potential promoter and regulatory elements of *HEY1*. When using this approach in existing literature, it is suggested that the *HEY1* promotor region is hypermethylated in health tissue (Tsung *et al*, 2017). In this data set, a region was identified as being hypermethylated (Chr8: 79,770,443 and Chr8: 79,770,914) in the reference genome in a bottle dataset which could suggest that this is a promotor or regulatory region. However, biological conclusions cannot be drawn from this as the analysis relies of reference data sets rather than patient matched samples. In future studies, we will use patient specific germline samples as the reference when assessing variation in base modifications.

Long-read single-nuclei sequencing

The quality control metrics obtained following nuclei isolation, cDNA amplification and ONT library preparation indicate that the experimental workflow was largely successful across all samples. As the aim of the protocol is to isolate nuclei and not intact cells, low viability is to be expected with higher values indicating incomplete lysis and potential contamination of whole cells. Nuclei viability was ≤0.1% in all cases, which is consistent with effective nuclei dissociation. The nuclei concentrations were sufficient for GEM-X loading, and the

estimated capture efficiency of 80% suggests effective encapsulation of nuclei and minimal loss during processing.

Tapestation analysis following cDNA amplification demonstrated fragment size predominantly within the expected range of 400-2000bp confirming successful generation of amplified cDNA suitable for long-read sequencing. In sample 1, a minor peak at 250bp was observed; however, its relatively low abundance to the dominant peak suggests the presence of a small proportion of shorter cDNA amplicons which are unlikely to adversely affect downstream processes.

Despite differences in cDNA input, all samples produced enough adapter ligated library for ONT sequencing. Final library yields ranged from 128 to 158 ng, exceeding the recommended input of 100 fmol based on assumed average fragment size of 1Kb. This demonstrates that the library preparation protocol is robust and capable of generating sequencing library even from samples with lower initial cDNA yields. Collectively, these results support the suitability of all three libraries for downstream long-read transcriptomics but highlights the impact of biological variability on library yield which is an important consideration when working with heterogenous tissue samples.

In this study, single-nuclei libraries from sample 23R2886 were sequenced at both shall and high read depth to directly compare the impact of sequencing depth on data quality and downstream analysis. The shallow read depth dataset captured a substantially high number of cells (33,707) but with a lot less reads per cell (2,2474). As expected, this resulted in a lower number of median gene counts and UMI counts. In contrast the high read depth data set captured fewer cells (6,340) but at a much greater depth with a mean of 23,751 reads per cell resulting in a higher gene count and UMI count per cell. However, even with the

increase read depth and higher number of expressed genes per cell, the clustering was consistent between the two datasets, with both datasets resulting in six clusters at a UMAP resolution of 0.2, which suggests that the cell populations can still be identified with shallowing sequencing.

For cell annotation, anchoring, which is a computational strategy to annotate cell identities was used. This method creates anchors across datasets based on the cells transcriptional similarity. The anchors allow label transfer of cell types from a well-annotated reference data set to a new dataset. Even though anchoring is a robust way to distinguish cell types, mistakes can be made and datasets can be assigned the wrong cell type labels and hence the importance of putting the dataset into a biological context to check the accuracy. Confidence scores were generated when assigning cell types, which when low, could lead to misassignment.

The three single-cell datasets showed that T-cells were the most abundant cell population when compared to reference atlas 2 and when the best prediction annotation was generated from the highest annotation score for individual cells. MCS are known to have sparse immune cell infiltration, because of having high levels of extracellular matrix (EM) which can act as a physical barrier to immune infiltration (Files, *et* al, 2023; Walter *et al,* 2023).

It is likely that the abundant T-cell population in this samples is misclassification of cell types. Firstly, when comparing individual cell annotations from atlas 1 and atlas 2 a high proportion of cells annotated as T-cells by atlas 2 are annotated as Mesenchymal cells by atlas 1.

Secondly, in these two single-cell datasets, when comparing the 15 topmost differentially expressed genes per cluster, the population initially annotated as T-cells did not express notable T-cell markers such as CD3. Instead, these cells were characterised by the expression of *VIM, SPARC, SPARCL1, FLT1, ITGA6, ACTB, ACTG1* and *TMSB4X.* These genes are commonly associated with endothelial and stromal cells. The transcriptional profile is inconsistent with immune cell phenotypes and indicates that the cluster represent vascular endothelial and stromal cell populations, suggesting that the apparent abundance of T-cells reflects annotation error rather than immune infiltration in the sample.

As these are novel data sets, it could be that the reference datasets, Human Limb Embryo Atlas (Zhang *et al*, 2024) and Endochondral Ossification atlas (Lawrence *et al,* 2025) are not fully representative of the cell types of present in the MSC samples. The human limb embryo atlas and the endochondral ossification atlas were used as MCS resembles embryonic limb mesenchyme and early chondrogenesis and represents the most biologically accurate annotation reference dataset. However, manual labelling of cells could help identify novel cell clusters and correct low confidence anchored annotation.

Across all three samples, the *HEY1::NCOA2* fusion was successfully detected in a subset of nuclei, with counts ranging from 151 to 328 fusion-positive cells. In all samples, the fusion positive nuclei were dispersed across the UMAP, rather than forming distinct cluster which suggests that the fusion is not restrict to a defined cell population. This indicates that the fusion may occur across multiple cells states and is not restricted to the primitive cell population. However, following cell annotation, the fusion positive cells were distributed across mesenchymal, chondroprogenitor cells which is consistent with previous IHC studies (Aigner *et al,* 2000). The presence of fusion positive cells in the immune cell population,

again suggests annotation error. When looking at cell annotation using atlas 1, the fusion was only detected in a very sparse number (<5%) of cells annotated as chondrocytes which could support the hypothesis that the fusion is lost in the differentiated component of the tumour.

Overall, the fusion positive cells occurred mostly in the mesenchymal and chondroprogenitor cell populations, with mesenchymal cells comprising the largest proportion of fusion-positive cells across the samples (ranging from 35% to 52%).

4.5.1 Main Conclusions

- **Long-read vs Short-read sequencing:** Both short-read and long-read WGS reliably detected the fusion gene *HEY1::NCOA2* fusion genes in MCS samples.
- **Other variants:** Variants in *CHEK2, LEF1, and CCDC6* may contribute to tumour biology although their role in MCS is not yet established.
- **LR epigenetic analysis:** 5mC and 5hmC base modifications can be detected by LR, but patient-matched germline controls are needed for biological interpretation.
- **Single-nuclei sequencing:** protocol optimisation showed that this method is robust in library preparation and sequencing depth affects the gene/UMI counts but not the overall cell population clustering. The *HEY1:NCOA2* fusion can be detected in this dataset but occurs in multiple cell population, primarily mesenchymal and chondroprogenitor cells. Abundant T-cell populations are likely due to annotation errors, emphasizing the importance of manual annotation in cell type application.

Chapter 5 Conclusions

5.1 Overall conclusions

Overall, the findings demonstrate that modern tissue handling and preservation techniques can reliably support downstream molecular analysis. Both Hibernate-A and TissueReady were effective at maintaining tissue morphology and DNA integrity when samples are processed promptly, with minimal differences in histological quality. For preserving tissue morphology of frozen tissue, PrestoCHILL offers a rapid and safe alternative to using liquid nitrogen vapour. Tissue morphology was preserved without ice crystal artefact and DNA quality and quantity was suitable for molecular analysis. While both Monarch and Zymo DNA extractions kits produced DNA of comparable purity and concentration, the Monarch kit is better suited for applications requiring long DNA fragments. Together, these results provided practical guidance for selection of tissue preservation methods and extraction protocols depending on the downstream applications.

Methylation-base classification using the ROBIN analytical pipeline showed strong performance and was successful in classifying post-mortem cases despite the challenges of working with post-mortem tissue. Misclassification of two cases by the brain tumour classifiers demonstrates the importance of using pan-cancer classifiers particularly with rare tumour entities, the pan-cancer classifier was able to correctly classify cases, the brain classifiers could not. In addition, ROBIN reliably identified key copy number alterations, included CNV's, amplifications and deletions which are characteristics of specific tumour entities. However, the absence of adaptive sampling, limited the overall sequencing coverage and consequently limited fusion detection. While long-read sequencing showed concordance with short-read stand-of-care methods, this platform failed to detect clinically

relevant variants with low allele frequency scores. This demonstrates the need for high sequencing depth to achieve comprehensive structural variant detection.

Finally, in this thesis, pipelines were established for comparative DNA and RNA ONT long-read sequencing of MCS at bulk and single cell level. Long-read whole genome sequencing was successful at detecting the characteristic fusion of mesenchymal chondrosarcoma, HEY1::NCOA2. Additional variants in CHEK2, LEF1, and CCDC6 were detected and may play a role in tumour biology, although their relevance to MCS would need further investigation. Optimisation of single-nuclei protocols demonstrated robust library preparation. The HEY1::NCOA2 fusions was detected across multiple cell populations, predominantly mesenchymal and chondroprogenitor cells, while the apparent abundance of T-cell likely reflects mislabelling, demonstrating the importance of manual cell-type annotation.

Together, these findings support the foundations of a rapid, non-toxic and cost-effective pathway using Oxford Nanopore long-read sequencing platforms for near-patient precision oncology for people with rare cancers.

5.2 Future directions

ONT sequencing with methylation-based classification has been successful on frozen tissue from post-mortem cases. More recently, research has shown that ONT can sequence DNA derived from FFPE tissue for methylation based classification. With the PM cohort described in chapter 4, FFPE tumour blocks would be selected and ONT WGS performed from DNA derived from these blocks. The sequencing data would be run through the ROBIN pipeline to explore if the classifier results are comparable with the results from frozen tissue and EPIC data if available.

The Tess Jowell BRAIN MATRIX study aims to develop infrastructure to provide rapid integrated histomolecular diagnostics of gliomas. To date, patients registered to BRAIN MATRIX have had short-read whole genome sequencing via the standard-of-care pathway. For some recruiting centres, including Oxford, there will be a long-read sub-study. Oxford BRAIN MATRIX patients will form a cohort of cases that will have prospective ONT sequencing and methylation classification using ROBIN.

For some rare cancers in children and young adults, it can take a long time to reach a diagnosis following extensive testing. In some cases, a definite diagnosis might not be reached. Future plans include using ONT and ROBIN for a rapid tissue diagnostic service for rare solid cancers in child, teenagers and young adults. For a pilot study, cases with molecular data such as short-read WGS or EPIC available would be selected and sequenced with ONT, and data analysed through the ROBIN pipeline methylation classifiers. With this project, a protocol will be optimised to test if the methylation classifiers can work on circulating DNA derived from liquid biopsies.

This thesis contributed to the first bulk and single cell ONT long-read datasets of MCS. However, the cohort size was small (n=10) and sequencing coverage was ranged from 13X to 30X. As described in chapter 3, to achieve comprehensive structural variant detection, high sequencing depth is required. In addition, one aim of this chapter was to use long-read sequencing (ONT) to investigate the epigenetic landscape of MCS and identify regions of hypermethylation or hypomethylation. Whilst 5mC and 5hmC base modifications were detected by long-read sequencing, biological relevance could not be interpreted due to using Genome in a bottle as a reference data set. To assess the potential clinical relevance of mutation detected and epigenetic modifications, future work will include long-read whole genome sequencing on a subset of cases in which patient matched germline samples are available. This will allow for more accurate biological interpretation between disease causing base modifications and the patients normal epigenetic modifications. For this, guidelines from genomics England of sequencing tumour samples to a depth of 60x and germline samples to a depth of 30x will be followed.

Finally in chapter 4, long-read single-nuclei protocol was optimised. However, the bioinformatic analysis pipeline needs improvement. In future work, removal of ambient RNA will be included in the pipeline. Ambient RNA can contaminant the sample by be encapsulated in a droplet forming a GEM, this can lead to misinterpretation of results and incorrect annotation of cells. Further to this, the pipeline will include separating spliced vs non-spiced transcripts. Un-spliced transcripts represent early transcription phase, whereas spliced RNA represents mature transcripts, enabling an understanding of cell state and differentiation trajectories. Finally, manual annotation will be applied to these datasets. These datasets are novel and reference datasets are not fully representative of cell types present in these samples. Manual annotation will allow for more accurate annotation. This will be performed

using Seurat to identified genes that are differentially expressed across the clusters and comparing cluster specific markers to known cell type markers to identify cell populations.

Bibliography

10x Genomics. (2024). *Chromium GEM-X Single Cell 5' Reagent Kits v3*. www.10xgenomics.com/trademarks.

Adams, J. U. (2008). The Human Genome project set out to sequence all of the 3 billion nucleotides in the human genome. Exactly how was this daunting task done with such incredible speed and accuracy? *Nature Education*, *1*(1), 193.

Aigner, T., Loos, S., Mü, S., Sandell, L. J., Krishnan Unni, K., & Kirchner, T. (2000). Cell Differentiation and Matrix Gene Expression in Mesenchymal Chondrosarcomas. In *Am J Pathol* (Vol. 156).

Antoni, L., Sodha, N., Collins, I., & Garrett, M. D. (2007). CHK2 kinase: Cancer susceptibility and cancer therapy - Two sides of the same coin? In *Nature Reviews Cancer* (Vol. 7, Issue 12, pp. 925–936). https://doi.org/10.1038/nrc2251

Atelerix. (2020.). *TissueReady™ Handbook TR07;0.1.1*. Retrieved January 13, 2026, from https://143938571.fs1.hubspotusercontent-eu1.net/hubfs/143938571/Protocols%20(2025%20branding)/TissueReady%20Protocol%20TR07_0.1.1.pdf

Aterlerix. (2021). *Storage and transport of human cells and room temperature* . https://www.conferenceharvester.com//uploads/harvester/exhibitors/134/BPFLSJUN-PDF-426569-390030-1-PDF(1).pdf#:~:text=Page%208,%E2%80%A2%20PRIMARY%20SKIN%20BIOPSIES

Aterlerix. (2025). *Preserving cancer tissue at room temperature using TissueReady* . https://143938571.fs1.hubspotusercontent-eu1.net/hubfs/143938571/Brochures%20and%20Data%20Packs/Copy%20of%20Atelerix%20Datapack%20v0.14.pdf

Barrachina, M., & Ferrer, I. (n.d.). *DNA Methylation of Alzheimer Disease and Tauopathy-Related Genes in Postmortem Brain*. www.jneuropath.com

Bender, E. (2018). Getting cancer drugs into the brain. *Nature*, *561*(7724), S46–S47. https://doi.org/10.1038/d41586-018-06707-4

Benfatto, S., Sill, M., Jones, D. T. W., Pfister, S. M., Sahm, F., von Deimling, A., Capper, D., & Hovestadt, V. (2025). Explainable artificial intelligence of DNA methylation-based brain tumor diagnostics. *Nature Communications* , *16*(1). https://doi.org/10.1038/s41467-025-57078-0

Biegel, J. A., Zhou, J. Y., Rorke, L. B., Stenstrom, C., Wainwright, L. M., & Fogelgren, B. (1999). Germ-line and acquired mutations of INI1 in atypical teratoid and rhabdoid tumors. *Cancer Research*, *59*(1), 74–79.

Brandin, E., Branton, D., & Deamer, D. W. (1996). Characterization of Individual Polynucleotide Molecules Using a Membrane Channel. In *Source* (Vol. 93, Issue 24).

Brewer, G. J., & Price, P. J. (1996). Viable cultured neurons in ambient carbon dioxide and hibernation storage for a month. *NeuroReport*, *7*(9), 1509–1512.

Buas, M. F., Kabak, S., & Kadesch, T. (2010). The Notch Effector Hey1 Associates with Myogenic Target Genes to Repress Myogenesis. *Journal of Biological Chemistry*, *285*(2), 1249–1258. https://doi.org/10.1074/jbc.M109.046441

Buckley, A. R., Ideker, T., Carter, H., & Schork, N. J. (2019). Rare variant phasing using paired tumor:normal sequence data. *BMC Bioinformatics*, *20*(1). https://doi.org/10.1186/s12859-019-2753-1

Cairncross, J. G., & Macdonald, D. R. (1988). Successful chemotherapy for recurrent malignant oligodendroglioma. *Annals of Neurology*, *23*(4), 360–364. https://doi.org/10.1002/ana.410230408

Cairncross, J. G., Ueki, K., Zlatescu, M. C., Lisle, D. K., Finkelstein, D. M., Hammond, R. R., Silver, J. S., Stark, P. C., Macdonald, D. R., Ino, Y., Ramsay, D. A., & Louis, D. N. (1998). Specific Genetic Predictors of Chemotherapeutic Response and Survival in Patients With Anaplastic Oligodendrogliomas. *Journal of the National Cancer Institute*, *90*(19), 1473–1479. https://doi.org/10.1093/jnci/90.19.1473

Capper, D., Jones, D. T. W., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., Koelsche, C., Sahm, F., Chavez, L., Reuss, D. E., Kratz, A., Wefers, A. K., Huang, K., Pajtler, K. W., Schweizer, L., Stichel, D., Olar, A., Engel, N. W., Lindenberg, K., … Pfister, S. M. (2018a). DNA methylation-based classification of central nervous system tumours. *Nature*, *555*(7697), 469–474. https://doi.org/10.1038/nature26000

Capper, D., Stichel, D., Sahm, F., Jones, D. T. W., Schrimpf, D., Sill, M., Schmid, S., Hovestadt, V., Reuss, D. E., Koelsche, C., Reinhardt, A., Wefers, A. K., Huang, K., Sievers, P., Ebrahimi, A., Schöler, A., Teichmann, D., Koch, A., Hänggi, D., … von Deimling, A. (2018b). Practical implementation of DNA methylation and copy-number-based CNS tumor diagnostics: the Heidelberg experience. *Acta Neuropathologica*, *136*(2), 181–210. https://doi.org/10.1007/s00401-018-1879-y

Cerrato, A., Merolla, F., Morra, F., & Celetti, A. (2018). CCDC6: the identity of a protein known to be partner in fusion. In *International Journal of Cancer* (Vol. 142, Issue 7, pp. 1300–1308). Wiley-Liss Inc. https://doi.org/10.1002/ijc.31106

Coakham, H. B., Garson, J. A., Browneli, B., & Kemshead, J. T. (1984). Monoclonal antibodies as reagents for brain tumour diagnosis: a review. *Journal of the Royal Society of Medicine*, *77*.

Collins, V. P. (2004). Brain tumours: Classification and genes. *Neurology in Practice*, *75*(2). https://doi.org/10.1136/jnnp.2004.040337

Colquitt, B. M., Allen, W. E., Barnea, G., & Lomvardas, S. (2013). Alteration of genic 5-hydroxymethylcytosine patterning in olfactory neurons correlates with changes in gene expression and cell identity. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(36), 14682–14687. https://doi.org/10.1073/pnas.1302759110

Coons, Albert. H., Creech, Hugh. J., & Jones, R. N. (1941). Immunological Properties of an Antibody Containing a Fluorescent Group. *Experimental Biology and Medicine* , *47*(2).

Cortés-Ciriano, I., Elrick, H., Sauer, C., Valle-Inclan, J. E., Trevers, K., Tanguy, M., Zumalave, S., de Noon, S., Muyas, F., Cascao, R., Afonso, A., Amary, F., Tirabosco, R., Giess, A., Freeman, T., Sosinsky, A., Piculell, K., Miller, D., Faria, C., … Flanagan, A. (2024). *SAVANA: reliable analysis of somatic structural variants and copy number aberrations in clinical samples using long-read sequencing*. https://doi.org/10.21203/rs.3.rs-4870639/v1

CRUK. (2025, July 31). *Brain, other CNS and intracranial tumours statistics*. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/brain-other-cns-and-intracranial-tumours/incidence

Cushing, H., & Bailey, P. (1927). A Classification of the Tumors of the Glioma Group on a Histogenetic Basis with a Correlated Study of Prognosis. *Archives of Neurology and Psychology* , *17*(4), 570.

Cybulski, C., Górski, B., Huzarski, T., Masojc´,Masojc´, B., Mierzejewski, M., de ¸bniak, T., Teodorczyk, U., Byrski, T., Gronwald, J., Matyjasik, J., Złowocka, E., Lenner, M., Grabowska, E., Nej, K., Castaneda, J., Me ¸drek, K., Szyman´ska, A. S., Szyman´ska, J. S., Kurzawski, G., Lubin´ski, J. L. (2004). CHEK2 Is a Multiorgan Cancer Susceptibility Gene. In *Am. J. Hum. Genet* (Vol. 75).

de Ruijter, T. C., de Hoon, J. P., Slaats, J., de Vries, B., Janssen, M. J., van Wezel, T., Aarts, M. J., van Engeland, M., Tjan-Heijnen, V. C., van Neste, L., & Veeck, J. (2015). Formalin-fixed, paraffin-embedded (FFPE) tissue epigenomics using Infinium HumanMethylation450 BeadChip assays. *Laboratory Investigation*, *95*(7), 833–842. https://doi.org/10.1038/labinvest.2015.53

Deacon, S., Cahyani, I., & Loose, M. (2024). *ROBIN: A unified nanopore-based sequencing assay integrating real-time, intraoperative methylome classification and next-day comprehensive molecular brain tumour profiling for ultra-rapid tumour diagnostics V.2* . https://www.protocols.io/view/robin-a-unified-nanopore-based-sequencing-assay-in-bp2l6xepklqe/v2?version_warning=no

Deacon, S., Cahyani, I., Holmes, N., Fox, G., Munro, R., Wibowo, S., Murray, T., Mason, H., Housley, M., Martin, D., Sharif, A., Patel, A., Goldspring, R., Brandner, S., Sahm, F., Smith, S., Paine, S., & Loose, M. (2025). ROBIN: A unified nanopore-based assay integrating

intraoperative methylome classification and next-day comprehensive profiling for ultra-rapid tumor diagnosis. *Neuro-Oncology*. https://doi.org/10.1093/neuonc/noaf103

Deamer, D., Akeson, M., & Branton, D. (2016). Three decades of nanopore sequencing. In *Nature Biotechnology* (Vol. 34, Issue 5, pp. 518–524). Nature Publishing Group. https://doi.org/10.1038/nbt.3423

Diaz-Perez, J. A., Nielsen, G. P., Antonescu, C., Taylor, M. S., Lozano-Calderon, S. A., & Rosenberg, A. E. (2019). EWSR1/FUS-NFATc2 rearranged round cell sarcoma: clinicopathological series of 4 cases and literature review. *Human Pathology*, *90*, 45–53. https://doi.org/10.1016/j.humpath.2019.05.001

Dudzisz-Śledź, M., Kondracka, M., Rudzińska, M., Zając, A. E., Firlej, W., Sulejczak, D., Borkowska, A., Szostakowski, B., Szumera-Ciećkiewicz, A., Piątkowski, J., Rutkowski, P., & Czarnecka, A. M. (2023). Mesenchymal Chondrosarcoma from Diagnosis to Clinical Trials. In *Cancers* (Vol. 15, Issue 18). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/cancers15184581

el Beaino, M., Roszik, J., Livingston, J. A., Wang, W.-L., Lazar, A. J., Amini, B., Subbiah, V., Lewis, V., & Conley, A. P. (2018). Mesenchymal Chondrosarcoma: a Review with Emphasis on its Fusion-Driven Biology. *Current Oncology Reports*, *20*(5), 37. https://doi.org/10.1007/s11912-018-0668-z

El-Gamal, M. I., Mewafi, N. H., Abdelmotteleb, N. E., Emara, M. A., Tarazi, H., Sbenati, R. M., Madkour, M. M., Zaraei, S. O., Shahin, A. I., & Anbar, H. S. (2021). A review of her4 (Erbb4) kinase, its impact on cancer, and its inhibitors. In *Molecules* (Vol. 26, Issue 23). MDPI. https://doi.org/10.3390/molecules26237376

Ellison, D. W., & Taylor, M. D. (2021). *Medulloblastoma In: WHO Classification of Tumours Editorial Board. Central nervous system tumours* (P. Wesseling & S. Pfister, Eds.; 5th ed., Vol. 6). International Agency for Research on Cancer . https://tumourclassification.iarc.who.int/chapters/33

Epi2me-labs. (2021). *wf-single-cell Nextflow workflow (version 3.3.2)*. https://github.com/Epi2me-labs/wf-single-cell

Esteller, M., Garcia-Foncillas, J., Andion, E., Goodman, S., Hidalgo, O., Vanacloscha, V., Baylin, S., & Herman, J. (2000). Epigenetic lesions causing genetic lesions in human cancer promoter hypermethylation of DNA repair genes. *European Journal of Cancer*, *36*, 2294–2300.

Euskirchen, P., Bielle, F., Labreche, K., Kloosterman, W. P., Rosenberg, S., Daniau, M., Schmitt, C., Masliah-Planchon, J., Bourdeaut, F., Dehais, C., Marie, Y., Delattre, J. Y., & Idbaih, A. (2017). Same-day genomic and epigenomic diagnosis of brain tumors using real-time nanopore sequencing. *Acta Neuropathologica*, *134*(5), 691–703. https://doi.org/10.1007/s00401-017-1743-5

Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., di Tommaso, P., & Nahnsen, S. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*, *38*(3), 276–278. https://doi.org/10.1038/s41587-020-0439-x

Fanburgh-Smith, J., Ladanyi, M., & de Pinieux, G. (2020). *Mesenchymal chondrosarcoma In: WHO Classification of Tumours Editorial Board. Soft tissue and bone tumours* (J. Bovée, Ed.; 5th ed., Vol. 3). International Agency for Research on Cancer. https://tumourclassification.iarc.who.int/chapters/33

Flies, D. B., Langermann, S., Jensen, C., Karsdal, M. A., & Willumsen, N. (2023). Regulation of tumor immunity and immunotherapy by the tumor collagen extracellular matrix. In *Frontiers in Immunology* (Vol. 14). Frontiers Media SA. https://doi.org/10.3389/fimmu.2023.1199513

Frezza, A. M., Cesari, M., Baumhoer, D., Biau, D., Bielack, S., Campanacci, D. A., Casanova, J., Esler, C., Ferrari, S., Funovics, P. T., Gerrand, C., Grimer, R., Gronchi, A., Haffner, N., Hecker-Nolting, S., Höller, S., Jeys, L., Jutte, P., Leithner, A., Whelan, J. (2015). Mesenchymal chondrosarcoma: Prognostic factors and outcome in 113 patients. A European Musculoskeletal Oncology Society study. *European Journal of Cancer*, *51*(3), 374–381. https://doi.org/10.1016/j.ejca.2014.11.007

Gallagher, S. (1998). Quantitation of Nucleic Acids with Absorption Spectroscopy. *Current Protocols in Protein Science*, *13*(1). https://doi.org/10.1002/0471140864.psa04ks13

Goldring, M. B., Tsuchimochi, K., & Ijiri, K. (2006). The control of chondrogenesis. In *Journal of Cellular Biochemistry* (Vol. 97, Issue 1, pp. 33–44). https://doi.org/10.1002/jcb.20652

Greenberg, M. V. C., & Bourc'his D. (2019). The diverse roles of DNA methylation in mammalian development and disease. *Nature Reviews Molecular Cell Biology*, *20*, 590–607.

Grizzle, W. E., Bell, W. C., & Sexton, K. C. (2011). Issues in collecting, processing and storing human tissues and associated information to support biomedical research. *Cancer Biomarkers*, *9*(1–6), 531–549. https://doi.org/10.3233/CBM-2011-0183

Gross, J. A., Nagy, C., Lin, L., Bonneil, É., Maheu, M., Thibault, P., Mechawar, N., Jin, P., & Turecki, G. (2016). Global and site-specific changes in 5-methylcytosine and 5-hydroxymethylcytosine after extended post-mortem interval. *Frontiers in Genetics*, *7*(JUN). https://doi.org/10.3389/fgene.2016.00120

Han, Q., Liang, H., Cheng, P., Yang, H., & Zhao, P. (2020). Gross Total vs. Subtotal Resection on Survival Outcomes in Elderly Patients With High-Grade Glioma: A Systematic Review and Meta-Analysis. *Frontiers in Oncology*, *10*. https://doi.org/10.3389/fonc.2020.00151

Han, S.-H., Kim, K. W., Kim, S., & Youn, Y. C. (2018). Artificial Neural Network: Understanding the Basic Concepts without Mathematics. *Dementia and Neurocognitive Disorders*, *17*(3), 83. https://doi.org/10.12779/dnd.2018.17.3.83

Hanif, F., Muzaffar, K., Perveen, K., Malhi, Saima. M., & Simjee, Shabana. U. (2017). Glioblastoma multiforme: A review of its epidemiology and pathogenesis through clinical presentation and treatment. *Asian Pacific Journal of Cancer Prevention*, *18*(1), 3–9. https://doi.org/10.22034/APJCP.2017.18.1.3

Hartwig Medical Foundation. (2026). *ESVEE structural variant engine* . Https://Github.Com/Hartwigmedical/Hmftools/Blob/Master/Esvee/README.Md.

He, B., Yao, H., & Yi, C. (2024). Advances in the joint profiling technologies of 5mC and 5hmC. In *RSC Chemical Biology* (Vol. 5, Issue 6, pp. 500–507). Royal Society of Chemistry. https://doi.org/10.1039/d4cb00034j

Hegi, M. E., Diserens, A.-C., Gorlia, T., Hamou, M.-F., de Tribolet, N., Weller, M., Kros, J. M., Hainfellner, J. A., Mason, W., Mariani, L., Bromberg, E. C., Hau, P., Mirimanoff, R. O., Cairncross, J. G., Janzer, R. C., & Stupp, R. (2005). MGMT Gene Silencing and Benefit from Temozolomide in Glioblastoma. *The New England Journal of Medicine*, *352*, 997–1003. www.nejm.org

Heller, D., & Vingron, M. (2019). SVIM: structural variant identification using mapped long reads. *Bioinformatics*, *35*(17), 2907–2915. https://doi.org/10.1093/bioinformatics/btz041

Hermes, M., Geisler, H., Osswald, H., Riehle, R., & Kloor, D. (2008). Alterations in S-adenosylhomocysteine metabolism decrease O6-methylguanine DNA methyltransferase gene expression without affecting promoter methylation. *Biochemical Pharmacology*, *75*(11), 2100–2111. https://doi.org/10.1016/j.bcp.2008.02.031

Holliday, R., & Pugh, J. E. (1975). DNA Modification Mechanisms and Gene Activity during Development Developmental clocks may depend on the enzymic modification of specific bases in repeated DNA sequences. *Science*, *187*(4173), 226–232. https://www.science.org

Hu, Y., Ge, X., Xie, Q., Ma, R., & Tao, Q. (2025). Progress in the study of molecular markers in the prognosis assessment and recurrence patterns of glioblastoma. In *Cancer Biology and Therapy* (Vol. 26, Issue 1). Taylor and Francis Ltd. https://doi.org/10.1080/15384047.2025.2574179

Hunter, K., Alexander, A., Passerini, S., Rovner, A., & Garg, A. (2016). Extraskeletal mesenchymal chondrosarcoma arising in adductor magnus with metastatic foci. *BJR|case Reports*, *2*(1), 20150117. https://doi.org/10.1259/bjrcr.20150117

Jarmasz, J. S., Stirton, H., Davie, J. R., & del Bigio, M. R. (2019). DNA methylation and histone post-translational modification stability in post-mortem brain tissue. *Clinical Epigenetics*, *11*(1). https://doi.org/10.1186/s13148-018-0596-7

Jaudou, S., Tran, M. L., Vorimore, F., Fach, P., & Delannoy, S. (2022). Evaluation of high molecular weight DNA extraction methods for long-read sequencing of Shiga toxin-

producing Escherichia coli. *PLoS ONE*, *17*(7 July). https://doi.org/10.1371/journal.pone.0270751

Jiang, T., Liu, Y., Jiang, Y., Li, J., Gao, Y., Cui, Z., Liu, Y., Liu, B., & Wang, Y. (2020). Long-read-based human genomic structural variation detection with cuteSV. *Genome Biology*, *21*(1), 189. https://doi.org/10.1186/s13059-020-02107-y

Jones, S. (2004). *An overview of the basic helix-loop-helix proteins*. http://genomebiology.com/2004/5/6/226

Kersey, H. N., Acri, D. J., Dabin, L. C., Hartigan, K., Mustaklem, R., Park, J. H., & Kim, J. (2025). Comparative analysis of nuclei isolation methods for brain single-nucleus RNA sequencing. *BioRxiv : The Preprint Server for Biology*. https://doi.org/10.1101/2025.03.25.645306

Kishikawa, S., Kondo, A., Yao, T., & Saito, T. (2024). Case report: A mesenchymal chondrosarcoma with alternative HEY1::NCOA2 fusions in the sella turcica. *Pathology and Oncology Research*, *30*. https://doi.org/10.3389/pore.2024.1611730

Kleinschmidt-DeMasters, B. K., Flanagan, A., Inwards, C. Y., Baumhoer, D., Bouvier, C., & Hainfellner, J. A. (2021). *Mesenchymal chondrosarcoma In: WHO Classification of Tumours Editorial Board. Central nervous system tumours* (H. Ng & A. Lazar, Eds.; 5th ed., Vol. 6). International Agenecy for Research on Cancer . https://tumourclassification.iarc.who.int/chapters/45

Koelsche, C., Schrimpf, D., Stichel, D., Sill, M., Sahm, F., Reuss, D. E., Blattner, M., Worst, B., Heilig, C. E., Beck, K., Horak, P., Kreutzfeldt, S., Paff, E., Stark, S., Johann, P., Selt, F., Ecker, J., Sturm, D., Pajtler, K. W., … von Deimling, A. (2021). Sarcoma classification by DNA methylation profiling. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-020-20603-4

Kohler, G., & Milstein, C. (1975). Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature*, *256*, 495–497.

Kuschel, L. P., Hench, J., Frank, S., Hench, I. B., Girard, E., Blanluet, M., Masliah-Planchon, J., Misch, M., Onken, J., Czabanka, M., Karau, P., Ishaque, N., Hain, E. G., Heppner, F., Idbaih, A., Behr, N., Harms, C., Capper, D., & Euskirchen, P. (2021). *Robust methylation-based classification of brain tumors using nanopore sequencing*. https://doi.org/10.1101/2021.03.06.21252627

Lawrence, J. E. G., Woods, S., Roberts, K., Sumanaweera, D., Balogh, P., Li, T., Predeus, A. v., He, P., Polanski, K., Prigmore, E., Tuck, E., Mamanova, L., Zhou, D., Webb, S., Jardine, L., He, X., Barker, R. A., Haniffa, M., Flanagan, A. M., … Teichmann, S. A. (2025). Single-cell transcriptomics identifies chondrocyte differentiation dynamics in vivo and in vitro. *Developmental Cell*, *60*(22), 3066-3084.e8. https://doi.org/10.1016/j.devcel.2025.06.031

Laxmi, A., Gupta, P., & Gupta, J. (2019). CCDC6, a gene product in fusion with different protoncogenes, as a potential chemotherapeutic target. In *Cancer Biomarkers* (Vol. 24, Issue 4, pp. 383–393). IOS Press. https://doi.org/10.3233/CBM-181601

le Loarer, F., Szuhai, K., & Tirode, F. (2020). *Round cell sarcoma with EWSR1-non-ETS fusions In: WHO classification of tumours Editorial Board. Soft tissue and bone tumours* (J. A. Bridge, Ed.; 5th ed., Vol. 3). International Agency for Research on Cancer. https://tumourclassification.iarc.who.int/chapters/33

Leske, H., Camenisch Gross, U., Hofer, S., Neidert, M. C., Leske, S., Weller, M., Lehnick, D., & Rushing, E. J. (2023). MGMT methylation pattern of long-term and short-term survivors of glioblastoma reveals CpGs of the enhancer region to be of high prognostic value. *Acta Neuropathologica Communications*, *11*(1). https://doi.org/10.1186/s40478-023-01622-w

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, Y., & Tollefsbol, T. O. (2011). DNA Methylation Detection: Bisulfite Genomic Sequencing Analysis. In *Methods in Molecular Biology* (Vol. 791, pp. 11–21). Humana Press Inc. https://doi.org/10.1007/978-1-61779-316-5_2

Lin, G. L., & Monje, M. (2017). A protocol for rapid post-mortem cell culture of diffuse intrinsic pontine glioma (DIPG). *Journal of Visualized Experiments*, *2017*(121). https://doi.org/10.3791/55360

Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. In *Nature Reviews Genetics* (Vol. 21, Issue 10, pp. 597–614). Nature Research. https://doi.org/10.1038/s41576-020-0236-x

Louis, D. N., Perry, A., Burger, P., Ellison, D. W., Reifenberger, G., von Deimling, A., Aldape, K., Brat, D., Collins, V. P., Eberhart, C., Figarella-Branger, D., Fuller, G. N., Giangaspero, F., Giannini, C., Hawkins, C., Kleihues, P., Korshunov, A., Kros, J. M., Beatriz Lopes, M., Wesseling, P. (2014). International Society Of Neuropathology--Haarlem consensus guidelines for nervous system tumor classification and grading. *Brain Pathology (Zurich, Switzerland)*, *24*(5), 429–435. https://doi.org/10.1111/bpa.12171

Louis, D. N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W. K., Ohgaki, H., Wiestler, O. D., Kleihues, P., & Ellison, D. W. (2016). The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. In *Acta Neuropathologica* (Vol. 131, Issue 6, pp. 803–820). Springer Verlag. https://doi.org/10.1007/s00401-016-1545-1

Louis, D. N., Perry, A., Wesseling, P., Brat, D. J., Cree, I. A., Figarella-Branger, D., Hawkins, C., Ng, H. K., Pfister, S. M., Reifenberger, G., Soffietti, R., von Deimling, A., & Ellison, D. W. (2021). The 2021 WHO classification of tumors of the central nervous system: A summary. *Neuro-Oncology*, *23*(8), 1231–1251. https://doi.org/10.1093/neuonc/noab106

Lynch, M., & Conery, J. (2000). The Evolutionary Fate and Consequences of Duplicate Genes. *Science*, *290*(5494), 1151–1155. https://doi.org/DOI:10.1126/science.290.5494.115

Malley, D. S., Hamoudi, R. A., Kocialkowski, S., Pearson, D. M., Collins, V. P., & Ichimura, K. (2011). A distinct region of the MGMT CpG island critical for transcriptional regulation is preferentially methylated in glioblastoma cells and xenografts. *Acta Neuropathologica*, *121*(5), 651–661. https://doi.org/10.1007/s00401-011-0803-5

Malone, E. R., Oliva, M., Sabatini, P. J. B., Stockley, T. L., & Siu, L. L. (2020). Molecular profiling for precision cancer therapies. *Genome Medicine*, *12*(1). https://doi.org/10.1186/s13073-019-0703-1

Mantere, T., Kersten, S., & Hoischen, A. (2019). Long-read sequencing emerging in medical genetics. In *Frontiers in Genetics* (Vol. 10, Issue MAY). Frontiers Media S.A. https://doi.org/10.3389/fgene.2019.00426

Mckenzie, A. T., Thorn, E. L., Nnadi, O., Wróbel, B., Kendziorra, E., Farrell, K., Crary, J. F., & Brain, F. (2024). Cryopreservation of brain cell structure: a review Additional resources and electronic supplementary material: supplementary material. *Free Neuropathology*, *5*, 35. https://doi.org/10.17879/freeneuropathology-2024-5883

Meller, A., Nivon, L., & Branton, D. (2001). Voltage-driven DNA translocations through a nanopore. *Physical Review Letters*, *86*(15), 3435–3438. https://doi.org/10.1103/PhysRevLett.86.3435

Molinaro, A. M., Taylor, J. W., Wiencke, J. K., & Wrensch, M. R. (2019). Genetic and molecular epidemiology of adult diffuse glioma. In *Nature Reviews Neurology* (Vol. 15, Issue 7, pp. 405–417). Nature Publishing Group. https://doi.org/10.1038/s41582-019-0220-2

Moore, L. D., Le, T., & Fan, G. (2013). DNA methylation and its basic function. In *Neuropsychopharmacology* (Vol. 38, Issue 1, pp. 23–38). https://doi.org/10.1038/npp.2012.112

Morra, F., Merolla, F., Zito Marino, F., Catalano, R., Franco, R., Chieffi, P., & Celetti, A. (2021). The tumour suppressor CCDC6 is involved in ROS tolerance and neoplastic transformation by evading ferroptosis. *Heliyon*, *7*(11). https://doi.org/10.1016/j.heliyon.2021.e08399

Morra, F., Miro, C., Napolitano, V., Merolla, F., & Celetti, A. (2017). CCDC6 (coiled-coil domain containing 6). *Atlas of Genetics and Cytogenetics in Oncology and Haematology*, *4*. https://doi.org/10.4267/2042/62666

Moustakli, E., Christopoulos, P., Potiris, A., Zikopoulos, A., Mavrogianni, D., Karampas, G., Kathopoulis, N., Anagnostaki, I., Domali, E., Tzallas, A. T., Drakakis, P., & Stavros, S. (2025). Long-Read Sequencing and Structural Variant Detection: Unlocking the Hidden Genome in Rare Genetic Disorders. In *Diagnostics* (Vol. 15, Issue 14). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/diagnostics15141803

Müller, Y., Bühner, S., Fincke, V., Mauch-Mücke, K., Riemenschneider, M. J., Manea, S., Liesche-Starnecker, F., Hasselblatt, M., Dahlum, S., Boros, M., Siebert, R., Frühwald, M. C., & Johann, P. (2025). A rare case of atypical teratoid rhabdoid tumor (AT/RT) with homozygous SMARCB1 loss and one concurrent somatic heterozygous SMARCA4 variant. *Acta Neuropathologica Communications* , *13*(1). https://doi.org/10.1186/s40478-025-02129-2

Nakayama, R., Miura, Y., Ogino, J., Susa, M., Watanabe, I., Horiuchi, K., Anazawa, U., Toyama, Y., Morioka, H., Mukai, M., & Hasegawa, T. (2012). Detection of HEY1-NCOA2 fusion by fluorescence in-situ hybridization in formalin-fixed paraffin-embedded tissues as a possible diagnostic tool for mesenchymal chondrosarcoma. *Pathology International*, *62*(12), 823–826. https://doi.org/10.1111/pin.12022

New England Biosciences. (n.d.). *Product Manual for T3060 Quick Protocol Card for Tissuse Protocol for High Molecular Weight DNA (HMW DNA) Extraction from Tissue (NEB #T3060)*. https://www.neb.com/en-gb/protocols/protocol-for-high-molecular-weight-dna-hmw-dna-extraction-from-tissue?srsltid=AfmBOorjWf9-bgPC2LmNE5hXo4bwLOkFUOtjm8UytwkhBLCW6BrxMjWj

Newell-Price, J., Clark, A. J. L., & King, P. (2000). DNA Methylation and Silencing of Gene Expression. *Trends in Endocrinology and Metabolism*, *11*(4), 142–148.

Northcott, P. A., Taylor, M. D., Korshunov, A., Remke, M., Cho, Y. J., Clifford, S. C., Eberhart, C. G., Parsons, D. W., Rutkowski, S., Gajjar, A., Ellison, D. W., Lichter, P., Gilbertson, R. J., Pomeroy, S. L., Kool, M., & Pfister, S. M. (2012). Molecular subgroups of medulloblastoma: The current consensus. *Acta Neuropathologica*, *123*(4), 465–472. https://doi.org/10.1007/s00401-011-0922-z

Ohgaki, H., & Kleihues, P. (2005). Epidemiology and etiology of gliomas. *Acta Neuropathologica*, *109*(1), 93–108. https://doi.org/10.1007/s00401-005-0991-y

Olah, M., Patrick, E., Villani, A. C., Xu, J., White, C. C., Ryan, K. J., Piehowski, P., Kapasi, A., Nejad, P., Cimpean, M., Connor, S., Yung, C. J., Frangieh, M., McHenry, A., Elyaman, W., Petyuk, V., Schneider, J. A., Bennett, D. A., de Jager, P. L., & Bradshaw, E. M. (2018). A transcriptomic atlas of aged human microglia. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-02926-5

Oxford Nanopore Technologies. (2025). *modkit (v0.6.0)*. https://github.com/nanoporetech/modkit

Oxford Nanopore Technolgies. (2025). *Single-cell transcriptomics sequencing from 5' cDNA prepared with 10x Genomics using SQK-LSK114*. https://nanoporetech.com/document/ligation-sequencing-v14-single-cell-transcriptomics-with-5-cdna

Oxford Nanopore Technologies. (2025). *Ultra-Long DNA Sequencing Kit V14 (SQK-ULK114)*. https://nanoporetech.com/document/ultra-long-dna-sequencing-kit-sqk-ulk114

Panagopoulos, I., Gorunova, L., Bjerkehagen, B., Boye, K., & Heim, S. (2014). Chromosome aberrations and HEY1-NCOA2 fusion gene in a mesenchymal chondrosarcoma. *Oncology Reports*, *32*(1), 40–44. https://doi.org/10.3892/or.2014.3180

Patel, A., Dogan, H., Payne, A., Krause, E., Sievers, P., Schoebe, N., Schrimpf, D., Blume, C., Stichel, D., Holmes, N., Euskirchen, P., Hench, J., Frank, S., Rosenstiel-Goidts, V., Ratliff, M., Etminan, N., Unterberg, A., Dieterich, C., Herold-Mende, C., Sahm, F. (2022). Rapid-CNS2: rapid comprehensive adaptive nanopore-sequencing of CNS tumors, a proof-of-concept study. In *Acta Neuropathologica* (Vol. 143, Issue 5, pp. 609–612). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s00401-022-02415-6

Patel, A., Göbel, K., Ille, S., Hinz, F., Schoebe, N., Bogumil, H., Meyer, J., Brehm, M., Kardo, H., Schrimpf, D., Lomakin, A., Ritter, M., Göller, P., Kerbs, P., Pfeifer, L., Hamelmann, S., Blume, C., Ippen, F. M., Berghaus, N., Sahm, F. (2025). Prospective, multicenter validation of a platform for rapid molecular profiling of central nervous system tumors. *Nature Medicine*, *31*(5), 1567–1577. https://doi.org/10.1038/s41591-025-03562-5

Payne, A., Holmes, N., Clarke, T., Munro, R., Debebe, B. J., & Loose, M. (2021). Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nature Biotechnology*, *39*(4), 442–450. https://doi.org/10.1038/s41587-020-00746-x

Pedersen, B. S., & Quinlan, A. R. (2018). Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics*, *34*(5), 867–868. https://doi.org/10.1093/bioinformatics/btx699

Pidsley, R., Zotenko, E., Peters, T. J., Lawrence, M. G., Risbridger, G. P., Molloy, P., van Djik, S., Muhlhausler, B., Stirzaker, C., & Clark, S. J. (2016a). Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biology*, *17*(1). https://doi.org/10.1186/s13059-016-1066-1

Qi, W., Rosikiewicz, W., Yin, Z., Xu, B., Jiang, H., Wan, S., Fan, Y., Wu, G., & Wang, L. (2022). Genomic profiling identifies genes and pathways dysregulated by HEY1–NCOA2 fusion and shines a light on mesenchymal chondrosarcoma tumorigenesis. *Journal of Pathology*, *257*(5), 579–592. https://doi.org/10.1002/path.5899

Reifenberger, J., Reifenberger, G., Liu, L., David James, C., Wechsler, W., & Peter Collins, V. (1994). Molecular Genetic Analysis of Oligodendroglial Tumors Shows Preferential Allelic Deletions on 19q and 1p. *American Journal OfPathology*, *145*(5).

Remiszewski, P., Wąż, J., Falkowski, S., Rutkowski, P., & Czarnecka, A. M. (2025). Chemotherapy Strategies and Their Efficacy for Mesenchymal Chondrosarcoma. In *Current Oncology* (Vol. 32, Issue 11). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/curroncol32110615

Ren, X., Kang, B., & Zhang, Z. (2018). Understanding tumor ecosystems by single-cell sequencing: Promises and limitations. In *Genome Biology* (Vol. 19, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13059-018-1593-z

Romagnoli, S., Bartalucci, N., & Vannucchi, A. M. (2023). Resolving complex structural variants via nanopore sequencing. In *Frontiers in Genetics* (Vol. 14). Frontiers Media SA. https://doi.org/10.3389/fgene.2023.1213917

Roohani, S., Ehret, F., Perez, E., Capper, D., Jarosch, A., Flörcken, A., Märdian, S., Zips, D., & Kaul, D. (2022). Sarcoma classification by DNA methylation profiling in clinical everyday life: the Charité experience. *Clinical Epigenetics*, *14*(1). https://doi.org/10.1186/s13148-022-01365-w

Sahm, F., Brandner, S., Bertero, L., Capper, D., French, P. J., Figarella-Branger, D., Giangaspero, F., Haberler, C., Hegi, M. E., Kristensen, B. W., Kurian, K. M., Preusser, M., Tops, B. B. J., van den Bent, M., Wick, W., Reifenberger, G., & Wesseling, P. (2023). Molecular diagnostic tools for the World Health Organization (WHO) 2021 classification of gliomas, glioneuronal and neuronal tumors; An EANO guideline. In *Neuro-Oncology* (Vol. 25, Issue 10, pp. 1731–1749). Oxford University Press. https://doi.org/10.1093/neuonc/noad100

Sahm, F., Schrimpf, D., Jones, D. T. W., Meyer, J., Kratz, A., Reuss, D., Capper, D., Koelsche, C., Korshunov, A., Wiestler, B., Buchhalter, I., Milde, T., Selt, F., Sturm, D., Kool, M., Hummel, M., Bewerunge-Hudler, M., Mawrin, C., Schüller, U., von Deimling, A. (2016). Next-generation sequencing in routine brain tumor diagnostics enables an integrated diagnosis and identifies actionable targets. *Acta Neuropathologica*, *131*(6), 903–910. https://doi.org/10.1007/s00401-015-1519-8

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA Sequencing with Chain-Terminating Inhibitors. *PNAS*, *74*(12), 5463–5467.

Sankhe, C. S., Hall, L., & Kendall, G. C. (2025). Fusion oncogenes in rhabdomyosarcoma: model systems, mechanisms of tumorigenesis, and therapeutic implications. In *Frontiers in Oncology* (Vol. 15). Frontiers Media SA. https://doi.org/10.3389/fonc.2025.1570070

Santiago, L., Daniels, G., Wang, D., Deng, M., & Lee, P. (2017). Wnt signaling pathway protein LEF1 in cancer, as a biomarker for prognosis and a target for treatment. In *Am J Cancer Res* (Vol. 7, Issue 6). www.ajcr.us/

Santos, R., Lee, H., Williams, A., Baffour-Kyei, A., Lee, S. H., Troakes, C., Al-Chalabi, A., Breen, G., & Iacoangeli, A. (2025). Investigating the Performance of Oxford Nanopore Long-Read Sequencing with Respect to Illumina Microarrays and Short-Read Sequencing. *International Journal of Molecular Sciences*, *26*(10). https://doi.org/10.3390/ijms26104492

Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics*, *19*(R2). https://doi.org/10.1093/hmg/ddq416

Scheithauer, B. W. (2009). Development of the WHO classification of tumors of the central nervous system: A historical perspective. *Brain Pathology*, *19*(4), 551–564. https://doi.org/10.1111/j.1750-3639.2008.00192.x

Schloss, J. A. (2008). How to get genomes at one ten-thousandth the cost. *Nature Biotechnology* , *26*, 1113–1115.

Schwarze, K., Buchanan, J., Fermont, J. M., Dreau, H., Tilley, M. W., Taylor, J. M., Antoniou, P., Knight, S. J. L., Camps, C., Pentony, M. M., Kvikstad, E. M., Harris, S., Popitsch, N., Pagnamenta, A. T., Schuh, A., Taylor, J. C., & Wordsworth, S. (2019). The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom. *Genet Med*, *22*, 85–94. https://doi.org/10.1038/s41436

Schwarze, K., Buchanan, J., Taylor, J., & Wordsworth, S. (2018). Are whole Exome and whole Genome Sequencing Approaches Cost-Effective? A Systematic Review of the Literature. *Value in Health*, *21*, S100. https://doi.org/10.1016/j.jval.2018.04.677

Seager, M., Aisner, D. L., & Davies, K. D. (n.d.). *Oncogenic Gene Fusion Detection Using Anchored Multiplex Polymerase Chain Reaction Followed by Next Generation Sequencing HHS Public Access*.

Shabihkhani, M., Lucey, G. M., Wei, B., Mareninov, S., Lou, J. J., Vinters, H. v., Singer, E. J., Cloughesy, T. F., & Yong, W. H. (2014). The procurement, storage, and quality assurance of frozen blood and tissue biospecimens in pathology, biorepository, and biobank settings. *Clinical Biochemistry*, *47*(4–5), 258–266. https://doi.org/10.1016/j.clinbiochem.2014.01.002

Shakked, R. J., Geller, D. S., Gorlick, R., & Dorfman, H. D. (2012). Mesenchymal chondrosarcoma : Clinicopathologic study of 20 cases. *Archives of Pathology and Laboratory Medicine*, *136*(1), 61–75. https://doi.org/10.5858/arpa.2010-0362-OA

Shale, C., Cameron, D. L., Baber, J., Wong, M., Cowley, M. J., Papenfuss, A. T., Cuppen, E., & Priestley, P. (2022). Unscrambling cancer genomes via integrated analysis of structural variation and copy number. *Cell Genomics*, *2*(4). https://doi.org/10.1016/j.xgen.2022.100112

Shimizu, S., Sakamoto, K., Kudo, K., Morimoto, A., & Shioda, Y. (2023). Detection of BRAF V600E mutation in radiological Langerhans cell histiocytosis-associated neurodegenerative disease using droplet digital PCR analysis. *International Journal of Hematology*, *118*(1), 119–124. https://doi.org/10.1007/s12185-023-03588-w

Sill, M., Schrimpf, D., Patel, A., Sturm, D., Jäger, N., Sievers, P., Schweizer, L., Banan, R., Reuss, D., Suwala, A., Korshunov, A., Stichel, D., Wefers, A. K., Hau, A.-C., Boldt, H., Harter, P. N., Abdullaev, Z., Benhamida, J., Teichmann, D., … Sahm, F. (2025). Advancing CNS tumor diagnostics with expanded DNA methylation-based classification. *Cancer Cell*. https://doi.org/10.1016/j.ccell.2025.11.002

Simons, R. B., Karkala, F., Kukk, M. M., Adams, H. H. H., Kayser, M., & Vidaki, A. (2025). Comparative performance evaluation of bisulfite- and enzyme-based DNA conversion methods. *Clinical Epigenetics*, *17*(1). https://doi.org/10.1186/s13148-025-01855-7

Slyper, M., Porter, C. B. M., Ashenberg, O., Waldman, J., Drokhlyansky, E., Wakiro, I., Smillie, C., Smith-Rosario, G., Wu, J., Dionne, D., Vigneau, S., Jané-Valbuena, J., Tickle, T. L., Napolitano, S., Su, M. J., Patel, A. G., Karlstrom, A., Gritsch, S., Nomura, M., … Regev, A. (2020). A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. *Nature Medicine*, *26*(5), 792–802. https://doi.org/10.1038/s41591-020-0844-1

Smolka, M., Paulin, L. F., Grochowski, C. M., Horner, D. W., Mahmoud, M., Behera, S., Kalef-Ezra, E., Gandhi, M., Hong, K., Pehlivan, D., Scholz, S. W., Carvalho, C. M. B., Proukakis, C., & Sedlazeck, F. J. (2024). Detection of mosaic and population-level structural variants with Sniffles2. *Nature Biotechnology*, *42*(10), 1571–1580. https://doi.org/10.1038/s41587-023-02024-y

Szuhai, K., Ijszenga, M., de Jong, D., Karseladze, A., Tanke, H. J., & Hogendoorn, P. C. W. (2009). The NFATc2 Gene is involved in a novel cloned translocation in a ewing sarcoma variant that couples its function in immunology to oncology. *Clinical Cancer Research*, *15*(7), 2259–2268. https://doi.org/10.1158/1078-0432.CCR-08-2184

Tanaka, M., Homme, M., Teramura, Y., Kumegawa, K., Yamazaki, Y., Yamashita, K., Osato, M., Maruyama, R., & Nakamura, T. (2023). *HEY1-NCOA2 expression modulates chondrogenic differentiation and induces mesenchymal chondrosarcoma in mice*. https://doi.org/10.1172/jci

Thakkar, J. P., Dolecek, T. A., Horbinski, C., Ostrom, Q. T., Lightner, D. D., Barnholtz-Sloan, J. S., & Villano, J. L. (2014). Epidemiologic and molecular prognostic review of glioblastoma. *Cancer Epidemiology Biomarkers and Prevention*, *23*(10), 1985–1996. https://doi.org/10.1158/1055-9965.EPI-14-0275

Thavarajah, R., Mudimbaimannar, V. K., Elizabeth, J., Rao, U. K., & Ranganathan, K. (2012). Chemical and physical basics of routine formaldehyde fixation. In *Journal of Oral and Maxillofacial Pathology* (Vol. 16, Issue 3, pp. 400–405). https://doi.org/10.4103/0973-029X.102496

Torp, S. H., Solheim, O., & Skjulsvik, A. J. (2022). The WHO 2021 Classification of Central Nervous System tumours: a practical update on what neurosurgeons need to know—a minireview. *Acta Neurochirurgica*, *164*(9), 2453–2464. https://doi.org/10.1007/s00701-022-05301-y

Troskie, R. L., Jafrani, Y., Mercer, T. R., Ewing, A. D., Faulkner, G. J., & Cheetham, S. W. (2021). Long-read cDNA sequencing identifies functional pseudogenes in the human transcriptome. *Genome Biology*, *22*(1). https://doi.org/10.1186/s13059-021-02369-0

Tsung, A. J., Guda, M. R., Asuthkar, S., Labak, C. M., Purvis, I. J., Lu, Y., Jain, N., Bach, S. E., Prasad, D. V. R., & Velpula, K. K. (2017). Methylation regulates HEY1 expression in glioblastoma. In *Oncotarget* (Vol. 8, Issue 27). www.impactjournals.com/oncotarget/

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., & Thermes, C. (2018). The Third Revolution in Sequencing Technology. In *Trends in Genetics* (Vol. 34, Issue 9, pp. 666–681). Elsevier Ltd. https://doi.org/10.1016/j.tig.2018.05.008

Vermeulen, C., Pagès-Gallego, M., Kester, L., Kranendonk, M. E. G., Wesseling, P., Verburg, N., de Witt Hamer, P., Kooi, E. J., Dankmeijer, L., van der Lugt, J., van Baarsen, K., Hoving, E. W., Tops, B. B. J., & de Ridder, J. (2023). Ultra-fast deep-learned CNS tumour classification during surgery. *Nature*, *622*(7984), 842–849. https://doi.org/10.1038/s41586-023-06615-2

Waddington, C. H. (2012). The Epigenotype. *International Journal of Epidemiology*, *41*(1), 10–13. https://doi.org/10.1093/ije/dyr184

Walter, S. G., Knöll, P., Eysel, P., Quaas, A., Gaisendrees, C., Nißler, R., & Hieggelke, L. (2023). Molecular In-Depth Characterization of Chondrosarcoma for Current and Future Targeted Therapies. In *Cancers* (Vol. 15, Issue 9). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/cancers15092556

Wang, L., Motoi, T., Khanin, R., Olshen, A., Mertens, F., Bridge, J., Cin, P. D., Antonescu, C. R., Singer, S., Hameed, M., Bovee, J. V. M. G., Hogendoorn, P. C. W., Socci, N., & Ladanyi, M. (2012). Identification of a novel, recurrent HEY1-NCOA2 fusion in mesenchymal chondrosarcoma based on a genome-wide screen of exon-level expression data. *Genes Chromosomes and Cancer*, *51*(2), 127–139. https://doi.org/10.1002/gcc.20937

Wang, Y., Zhao, Y., Bollas, A., Wang, Y., & Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. In *Nature Biotechnology* (Vol. 39, Issue 11, pp. 1348–1365). Nature Research. https://doi.org/10.1038/s41587-021-01108-x

Wanis, H. A., Møller, H., Ashkan, K., & Davies, E. A. (2021). The incidence of major subtypes of primary brain tumors in adults in England 1995-2017. *Neuro-Oncology*, *23*(8), 1371–1382. https://doi.org/10.1093/neuonc/noab076

Wehrli, B. M., Huang, W., de Crombrugghe, B., Ayala, A. G., & Czerniak, B. (2003). Sox9, a master regulator of chondrogenesis, distinguishes mesenchymal chondrosarcoma from other small blue round cell tumors. *Human Pathology*, *34*(3), 263–269. https://doi.org/10.1053/hupa.2003.41

Weller, M., Wen, P. Y., Chang, S. M., Dirven, L., Lim, M., Monje, M., & Reifenberger, G. (2024). Glioma. *Nature Reviews Disease Primers*, *10*(1). https://doi.org/10.1038/s41572-024-00516-y

Xu, B., Rooper, L. M., Dermawan, J. K., Zhang, Y., Suurmeijer, A. J. H., Dickson, B. C., Demicco, E. G., & Antonescu, C. R. (2022). Mesenchymal chondrosarcoma of the head and neck with HEY1::NCOA2 fusion: A clinicopathologic and molecular study of 13 cases with emphasis on diagnostic pitfalls. *Genes Chromosomes and Cancer*, *61*(11), 670–677. https://doi.org/10.1002/gcc.23075

Yang, X., Tian, S., Fan, L., Niu, R., Yan, M., Chen, S., Zheng, M., & Zhang, S. (2022). Integrated regulation of chondrogenic differentiation in mesenchymal stem cells and differentiation of cancer cells. In *Cancer Cell International* (Vol. 22, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s12935-022-02598-8

Yu, W., Zhang, L., Wei, Q., & Shao, A. (2020). O6-Methylguanine-DNA Methyltransferase (MGMT): Challenges and New Opportunities in Glioma Chemotherapy. In *Frontiers in Oncology* (Vol. 9). Frontiers Media S.A. https://doi.org/10.3389/fonc.2019.01547

Yuan, D., Jugas, R., Pokorna, P., Sterba, J., Slaby, O., Schmid, S., Siewert, C., Osberg, B., Capper, D., Zeiner, P., Weber, K., Harter, P., Jabareen, N., Mackowiak, S., Ishaque, N., Eils, R., Lukassen, S., & Euskirchen, P. (2024). *crossNN: an explainable framework for cross-platform DNA methylation-based classification of cancer*. https://doi.org/10.1101/2024.01.22.24301523

Zhang, B., He, P., Lawrence, J. E. G., Wang, S., Tuck, E., Williams, B. A., Roberts, K., Kleshchevnikov, V., Mamanova, L., Bolt, L., Polanski, K., Li, T., Elmentaite, R., Fasouli, E. S., Prete, M., He, X., Yayon, N., Fu, Y., Yang, H., Teichmann, S. A. (2024). A human embryonic limb cell atlas resolved in space and time. *Nature*, *635*(8039), 668–678. https://doi.org/10.1038/s41586-023-06806-x

Zhang, H., Jain, C., & Aluru, S. (2020). A comprehensive evaluation of long read error correction methods. *BMC Genomics*, *21*. https://doi.org/10.1186/s12864-020-07227-0

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, *8*. https://doi.org/10.1038/ncomms14049

Zheng, Z., Li, S., Su, J., Leung, A. W.-S., Lam, T.-W., & Luo, R. (2022). Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *Nature Computational Science*, *2*(12), 797–803. https://doi.org/10.1038/s43588-022-00387-x

Zhou, Y., Chen, X., Chapman, J. S., & Barrett, M. T. (2024). Single nucleus DNA sequencing of flow sorted archived frozen and formalin fixed paraffin embedded solid tumors. *BMC Genomics*, *25*(1). https://doi.org/10.1186/s12864-024-10850-w

Zhou, J., Gu, L., Du, F., Li, C., Zhang, F., Zhang, X., Pang, J., Xie, B., Wang, X., Peng, J., & Jiang, Y. (2025). The global, regional, and national brain and CNS cancers burden and trends from 1990 to 2021. *Scientific Reports*, *15*(1). https://doi.org/10.1038/S41598-025-04636-7

Zhu, Z., Wu, X., Hu, Y., Bian, X., Wang, Y., & Zhu, Q. (2025). Alternative Splicing: Molecular Mechanisms, Biological Functions, Diseases, and Potential Therapeutic Targets. *MedComm*, *6*(12). https://doi.org/10.1002/mco2.70545

Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C. E., Alexander, N., Henaff, E., McIntyre, A. B. R., Chandramohan, D., Chen, F., Jaeger, E., Moshrefi, A., Pham, K., Stedman, W., Liang, T., Salit, M. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data*, *3*. https://doi.org/10.1038/sdata.2016.25

Zulch, K. J. (1979). *Histological typing of tumours of the central nervous system* (1st ed.). The World Health Organisation.

Zupanič Pajnič, I. (2025). Analysis of Human Degraded DNA in Forensic Genetics. In *Genes* (Vol. 16, Issue 11). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/genes16111375

Zymo. (n.d.). *Quick-DNA ™ Microprep Kit*. https://files.zymoresearch.com/protocols/_r1050_r1051_quick-rna_microprep_kit.pdf

Appendix

Appendix 1: Transfer of files from Windows to Ubuntu

```
crontab -e # use this to open the cron job file
*/10 * * * * rsync "location of input files" "/Location of output files"
>> Linux_home/log_from_time_chron_job.txt 2>&1
sudo service cron status
sudo service cron start
```

Appendix 2: ROBIN configuration file

```
[options]
# ROBIN needs to be aware of the targets being used in your experiment
bed_file =
storing_input_data_for_robin_PBA73002_pass_96b15088_0185d34f_0.bam

# The centreID is displayed in PDF reports generated by ROBIN
centreID = Neuropath_JR

# This is the port that the GUI will be displayed
port = 5678

# ROBIN needs to know where the reference file for aligning is
reference="/mnt/c/data/References files -USE THIS ONE/hg38.fa.gz"

# You can vary the number of threads used by ROBIN. For P2i devices this
should be 1
threads = 32

# The following options will be used in future improvements to ROBIN

# We record the basecall configuration in the options for future features
basecall_config =
dna_r10.4.1_e8.2_400bps_5khz_modbases_5hmc_5mc_cg_hac_prom.cfg
# We track the ideal experiment duration for future features
experiment_duration = 24
# We log the experiment kit being used for future features
kit = SQK-RAD114
```

Appendix 3: Oncoanalyser Nextflow Pipeline

The config file:

```
params {
genomes {
GRCh38_hmf {
fasta    =
"oncoanalyser_reference_samples/reference_data/2.0.0/20250604_114603/GRCh38
_masked_exclusions_alts_hlas.fasta"
fai    =
"oncoanalyser_reference_samples/reference_data/2.0.0/20250604_114603/GRCh38
_masked_exclusions_alts_hlas.fasta.fai"
dict    =
"oncoanalyser_reference_samples/reference_data/2.0.0/20250604_114603/GRCh38
_masked_exclusions_alts_hlas.fasta.dict"
img    =
"oncoanalyser_reference_samples/reference_data/2.0.0/20250604_114603/GRCh38
_masked_exclusions_alts_hlas.fasta.img"
bwamem2_index =
"oncoanalyser_reference_samples/reference_data/2.0.0/20250604_114603/bwa-
mem2_index_bwa-mem2_index-2.2.1/"
gridss_index =
"oncoanalyser_reference_samples/reference_data/2.0.0/20250604_114603/gridss
_index_gridss_index-2.13.2/"
}
}
}
```

The spreadsheet_for_oncoanalyser.csv looks like this:

```
group_id,subject_id,sample_id,sample_type,sequence_type,filetype,info,filep
ath
Sample_name,Sample_name,Sample_name_,tumor,dna,fastq,library_id:S1;lane:001
,Sample_name_R1_001.fastq.gz;Sample_name_R2_001.fastq.gz
```

Appendix 4: Long-read ONT sequencing pipeline

```
ALIGNING WITH MINIMAP2
echo "Beginning of Alignment with minimap2"

EXPLANATION OF SOFTWARE
### Samtools:
# Samtools merge will merge all the bam files you want to analyse together into a large file to analyse together
# Samtools index generates an index file named <filename.bai> which is used for quick access by other programmes
# Samtools index will make the output .bai file in the same directory as the .bam regardless of giving it a file path
# Samtools fastq will convert the bam file into a fastq format file, the -T means to include tags from the bam files
# in the outputted fastq file headers. -T MM,ML,mv are currently used, MM represents mapping quality,
# ML represents the mate read location and I don't know what mv is for
# the -@ 32 means to use 32 threads - makes things faster
#
### minimap2:
# -a indicates that the output should be in a SAM format.
# -x specifies the preset for the alignment
# lr:hq is a preset which means long read:high quality for fastand accurate mapping of nanopore data, (source: twitter not a manual)
# -t 24 means for minimap2 to use 24 threads
#
### mosdepth:
# --fast-mode means what it says, and --by 1000 means the average depth per 1000 base pairs is reported
#
### Order of command line:
# samtools merge -@ <number of threads> <name of merged output file> <bam file to merge 1> <bam file to merge 2> <...>
# samtools sort -@ <number of threads> <bam file> -o <output file name>
# minimap2 <mode and options> -t <number of threads> -d <minimap2 indexed reference genome> <input fastq> > <output file name>
# mosdepth --fast-mode --by <average depth for this number of bases> <prefix given to output files> <input file>
# the samtools fastq file as its being generated needs to be piped into minimap2 which is why those software are all on the same line
and don't state the input file, because the input ceoms from the pipe so the software knows to use it
#
### Ultimate outputs:
# ${FILE_PREFIX}.aligned.sorted.bam and
${FILE_PREFIX}.aligned.sorted.bam.bai

module purge
module load bluebear
module load bear-apps/2021b
module load SAMtools/1.15.1-GCC-11.2.0
# Align with minimap2
# Preprocessing
# samtools version samtools 1.15.1 using htslib 1.15.1

#you have made a large .bam file with samtools merge on the JR PC and so you can skip the first line of the code
```

```
# where the path bam files variable is set to the path and the example file
                            name you have given.
                            #samtools merge -@ 32
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.all.bam ${PATH_BAM_FILES}/bam_pass
   /*.bam        #will merge all the bam files in the bam_pass directory
                 samtools sort -@ 32 ${PATH_BAM_FILES} -o
 ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.all.sorted.bam    #!!! This is a
                new line I added from the original script
                            samtools index -@ 32
  ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.all.sorted.bam    #will make an
                             indexed .bai file
                                 # Aligning
                          # minimap2 version 2.27-r1193
                             samtools fastq -@ 4 -
    T MM,ML,mv ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.all.sorted.bam |
/rds/projects/b/beggsa-clinicalnanopore/minimap2/minimap2 -ax lr:hq -t 24 -
             y ${PATH_REF_GENOME_MMI} - | samtools sort -@ 4 -o
       ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.aligned.sorted.bam  -

                                 # postprocessing
                                samtools index -@ 32
        ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.aligned.sorted.bam
   #removing the files which you dont need/arent used in the next steps:
                                       rm
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.all.sorted.bam ${PATH_WORKING_DIRE
                  CTORY}/${FILE_PREFIX}.all.sorted.bam.bai
                                      date
                    echo "finished samtools and minimap2"

                              # Genome wide coverage
                             # mosdepth version 0.2.6
                          mkdir ${PATH_WORKING_DIRECTORY}/mosdepth
                 ${PATH_TO_SOFTWARE}/mosdepth --fast-mode --by 1000
                  ${PATH_WORKING_DIRECTORY}/mosdepth/${FILE_PREFIX}
        ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.aligned.sorted.bam
                                      date
                               echo "finished mosdepth"
                                     module purge
                                      date
                          echo "Finished alignment with minimap2"

              # Cramino for N50 number and other quality metrics
                          ${COMMAND_CONDA_ACTIVATE}/cramino
cramino -t 64 ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.aligned.sorted.bam >
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_cramino_quality_stats_outputs.txt
                          ${COMMAND_CONDA_DEACTIVATE}

                    echo "################################"

                          END OF ALIGNING WITH MINIMAP2

                       STRUCTURAL VARIANT CALLING PART1

                           EXPLANATION OF SOFTWARE
                                       #
                                  ### sniffles2:
   # --tandem-repeats will use tandem repeat annotation for the reference
                               genome that you provide
 # --mosaic will detect low-frequency SVs which don't occur in every cell
                                  ### cuteSV:
                                       #
```

```
#
### SVIM:
# The work directory is also the output directory of the VCF file
#
### Clair3:
# Requires the absolute paths to the directories
- eg. root/hom/workdir instead of just workdir
# --platform="ont" means oxford nanopore technologies , there are other
options for different DNA sequencing technologies
#
### Straglr-genotype:
#
### SpecHLA:
# you give the sample name but don't include the path to your work
directory as the files will be generated in specHLAs own work directory
#
### Modkit:
# --cpg means the programme counts CpG dinucleotides from a given reference
(--ref)
#
### Nanoplot:
# --alength Use aligned read lengths rather than sequenced length (bam
mode); --huge Input data is one very large file. ; -t 32 = 32 threads
# outputs: a statistical summary, a number of plots, a html summary file
#
### Savana:
# this is "savana to" which launches tumour only mode, the programme would
rather have a healthy sample and a tumour sample to compare against, but
you dont have this, so you just run tumour only mode
#
### Order of command line:
# sniffles -i <input bam file> -v <output vcf file name>  --tandem-repeats
<tandem repeat annotated reference genome> --threads <number of threads> --
mosaic
# cuteSV <input bam file> <reference fasta file> <output file name>
<specify a work directory>
# svim alignment <specify a work directory> <input bam file>
<reference fasta file>
# singularity exec clairs-to_latest.sif run_clair3.sh -T <input bam file> -
R <reference fasta file> -o <output path directory> -t <number of threads>
-p <platform/model of read analysis> --conda_prefix <path to
clair3 conda env>
# straglr-genotype --vcf <output vcf file name>--threads <number of
threads> --sample <sample prefix> --loci <path to wf_str_repeats_hg38.bed">
<input bam> <path to reference fasta file>
# ExtractHLAread.sh -s <sample name> -b <input sorted bam> -r hg38 -o
<output directory to store results>    #specHLA
# python3 long_read_typing.py -r <long read fastq generated
in previous step> -n <sample name> -o <output directory to store results> -
j <number of threads>    #specHLA
# modkit pileup <input sorted bam> <output bedMethyl file>
# modkit pileup <input sorted bam> <output CpG bedMethyl file> --cpg --ref
<path to reference genome>
# NanoPlot --bam <input sorted bam> -o <output file name> --alength -t 32 -
-huge
# savana savana to --tumour <input sorted bam>  --outdir <output file name>
--ref <path to reference genome>
#
### Ultimate outputs:
# ${FILE_PREFIX}.sniffles2.vcf | ${FILE_PREFIX}.cutesv.vcf
(${FILE_PREFIX}_cutesv_work is just a work directory for the software) |
```

```
${FILE_PREFIX}.svim.vcf (${FILE_PREFIX}_svim_work is just a work directory
for the software, I move the VCF file from here into the working directory
                                    for you)
  |  ${FILE_PREFIX}.clair3.indel.vcf.gz,  ${FILE_PREFIX}.clair3.snv.vcf.gz
  (clairs-to-${FILE_PREFIX} is the work directory for the software, I move
        the relevant VCF files into the working directory for you) |
${FILE_PREFIX}.straglr-genotype.vcf | ${FILE_PREFIX}_spechla_work, you need
 to go into this file and then into the folder with the sample name to see
       which VCF files from here will be useful as there are many. |
                  ${FILE_PREFIX}.modkit.allmods.bedMethyl,
                    ${FILE_PREFIX}.modkit.CpG.bedMethyl |
     ${FILE_PREFIX}_nanoplot (contains images and webpages of results)
##########################################################################
             #SV call with Sniffles2 - sniffles2 version 2.3.3
                    ${COMMAND_CONDA_ACTIVATE}/sniffles2
 sniffles -i ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.aligned.sorted.bam -v
  ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.sniffles2.vcf  --tandem-repeats
               ${PATH_REF_GENOME_TRF} --threads 32 --mosaic
                            ${COMMAND_CONDA_DEACTIVATE}
                                    date
                              echo "finished sniffles"

                # SV call with cuteSV - cuteSV version 2.0.2
                   rm -rf ${PATH_WORKING_DIRECTORY}/temp/
                      ${COMMAND_CONDA_ACTIVATE}/cutesv
         mkdir ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_cutesv_work
cuteSV ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.aligned.sorted.bam ${PATH_R
                                    EF_GENOME_FA}
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.cutesv.vcf ${PATH_WORKING_DIRECTOR
   Y}/${FILE_PREFIX}_cutesv_work/    # output file will be stored in work
                directory not inside cutesv_work directory
                            ${COMMAND_CONDA_DEACTIVATE}
                                    date
                              echo "finished cuteSV"

                  # SV call with SVIM - SVIM version 1.4.2
                                       #
                           ${COMMAND_CONDA_ACTIVATE}/svim
         mkdir ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_svim_work
                                    svim alignment
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_svim_work ${PATH_WORKING_DIRECTORY
         }/${FILE_PREFIX}.aligned.sorted.bam ${PATH_REF_GENOME_FA}
     mv ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_svim_work/variants.vcf
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.svim.vcf    # !!! new line added -
    this is to move the vcf made by svim and re-name it in line with the
                                    other vcf outputs
                            ${COMMAND_CONDA_DEACTIVATE}
                                    date
                               echo "finished svim"

                                    # Clair3
# Set Clair3 going somatic - no conda environment required - clair3 version
                                      1.0.7
                  singularity exec ${PATH_TO_SOFTWARE}/clairs-
                     to_latest.sif /opt/bin/run_clairs_to -T
        ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.aligned.sorted.bam -R
         ${PATH_REF_GENOME_FA} -o ${PATH_WORKING_DIRECTORY}/clairs-to-
              ${FILE_PREFIX}/ -t 32 -p ont_r10_dorado_hac_4khz --
                  conda_prefix /opt/micromamba/envs/clairs-to
```

```
mv ${PATH_WORKING_DIRECTORY}/clairs-to-${FILE_PREFIX}/indel.vcf.gz
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.clair3.indel.vcf.gz    #renaming
the files with the convention of the other files
mv ${PATH_WORKING_DIRECTORY}/clairs-to-${FILE_PREFIX}/snv.vcf.gz
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.clair3.snv.vcf.gz
date
echo "finished clair3"

# STR typing
# straglr-genotype version 1.4.2
${COMMAND_CONDA_ACTIVATE}/straglr-genotype
straglr-genotype --vcf ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.straglr-
genotype.vcf --threads 12 --sample ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}
--loci ${PATH_REF_GENOME_STRAGLR_BED}
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.aligned.sorted.bam ${PATH_REF_GENO
ME_FA}
${COMMAND_CONDA_DEACTIVATE}
date
echo "finished straglr"

#HLA typing
# specHLA version ???
# !!! specHLA conda is in a different location to the others, hence this
line
source /rds/homes/c/cavee/miniconda3/bin/activate /rds/projects/b/beggsa-
clinicalnanopore/software/SpecHLA2/SpecHLA/spechla_env
${PATH_TO_SOFTWARE}/SpecHLA2/SpecHLA/script/ExtractHLAread.sh -s
${FILE_PREFIX} -b
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.aligned.sorted.bam -r hg38 -o
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_spechla_work    # the -r flag only
takes input as hg38 or hg19, you don't need a file path to a reference
genome
python3 ${PATH_TO_SOFTWARE}/SpecHLA2/SpecHLA/script/long_read_typing.py -r
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_spechla_work/${FILE_PREFIX}_extrac
t.unpaired.fq.gz -n ${FILE_PREFIX} -o
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_spechla_work -j 32
${COMMAND_CONDA_DEACTIVATE}
date
echo "finished specHLA"

module purge; module load bluebear
#
# Run methylation profiling
module load bear-apps/2022b
module load SAMtools/1.17-GCC-12.2.0
#samtools version 1.15.1 using htslib 1.15.1
#modkit version 0.4.1
#
# Generate all mods files (e.g. not just CG motif but all mods) plus 5-mC
and 5-hmC BED files
#
${PATH_TO_SOFTWARE}/modkit/modkit pileup
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.aligned.sorted.bam ${PATH_WORKING_
DIRECTORY}/${FILE_PREFIX}.modkit.allmods.bedMethyl
${PATH_TO_SOFTWARE}/modkit/modkit pileup
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.aligned.sorted.bam ${PATH_WORKING_
DIRECTORY}/${FILE_PREFIX}.modkit.CpG.bedMethyl --cpg --ref
${PATH_REF_GENOME_FA}
# now make the IGV readable formatted files of these:
for strand in "+" "-"
do
```

```
for mod in "h" "m"
do
case ${strand} in
"+")
out_file=${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_allmods_${mod}_positive_I
GV.bedgraph
;;
"-")
out_file=${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_allmods_${mod}_negative_I
GV.bedgraph
;;
*)
echo "> not a strand"
exit 1
;;
esac
awk -v strand=${strand} -v mod=${mod} 'BEGIN{OFS="\t"} (($4==mod) &&
($6==strand)) {print $1,$2,$3,$11/100,$5}'
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.modkit.allmods.bedMethyl >
${out_file}
done
done

for strand in "+" "-"
do
for mod in "h" "m"
do
case ${strand} in
"+")
out_file=${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_CpG_${mod}_positive_IGV.b
edgraph
;;
"-")
out_file=${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_CpG_${mod}_negative_IGV.b
edgraph
;;
*)
echo "> not a strand"
exit 1
;;
esac
awk -v strand=${strand} -v mod=${mod} 'BEGIN{OFS="\t"} (($4==mod) &&
($6==strand)) {print $1,$2,$3,$11/100,$5}'
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.modkit.CpG.bedMethyl >
${out_file}
done
done

mkdir ${PATH_WORKING_DIRECTORY}/modkit_methylation_IGV_readable_files
mv
${PATH_WORKING_DIRECTORY}/*_positive_IGV.bedgraph ${PATH_WORKING_DIRECTORY}
/*_negative_IGV.bedgraph ${PATH_WORKING_DIRECTORY}/modkit_methylation_IGV_r
eadable_files

module purge
date
echo "finished modkit"

#
# Run Nanoplot on BAM file and generate read length data - NanoPlot version
1.42.0
```

```
#
${COMMAND_CONDA_ACTIVATE}/nanoplot
NanoPlot --bam
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.aligned.sorted.bam -o
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_nanoplot --alength -t 32 --huge
${COMMAND_CONDA_DEACTIVATE}
date
echo "finished nanoplot"
#
#run savana gene fusion caller - Savana version 1.3.1
#
${COMMAND_CONDA_ACTIVATE}/savana
savana to --tumour
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.aligned.sorted.bam  --
outdir ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_savana --ref
${PATH_REF_GENOME_FA}
${COMMAND_CONDA_DEACTIVATE}
date
echo "finished savana"

################ SOFTWARE WHICH USE R

# Run ASCAT

cd ${PATH_WORKING_DIRECTORY} # need to be in the actual directory for this
to work
#
module purge
module load bear-apps/2021b
module load R/4.2.0-foss-2021b
# R version 4.2.0
${COMMAND_CONDA_ACTIVATE}/alleleCounter
R CMD BATCH ${PATH_RUN_ASCAT_R_FILE}
${COMMAND_CONDA_DEACTIVATE}
# Tidying up the outputs from ASCAT
mv ${PATH_WORKING_DIRECTORY}/ascatploidy.txt
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_ascatploidy.txt
mv ${PATH_WORKING_DIRECTORY}/ascatpurity.txt
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_ascatpurity.txt
mkdir ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_ascat_outputs
mv ${PATH_WORKING_DIRECTORY}/*BAF*
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_ascat_outputs/
mv
${PATH_WORKING_DIRECTORY}/*.png ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_as
cat_outputs/
mv ${PATH_WORKING_DIRECTORY}/*alleleFrequencies*
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_ascat_outputs/
mv ${PATH_WORKING_DIRECTORY}/*.txt
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_ascat_outputs/
mv
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_ascat_outputs/${FILE_PREFIX}.segme
nts.txt ${PATH_WORKING_DIRECTORY}
mv
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.segments.txt ${PATH_WORKING_DIRECT
ORY}/${FILE_PREFIX}_ascat.segments.txt
mv
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_ascat_outputs/${FILE_PREFIX}_ascat
ploidy.txt ${PATH_WORKING_DIRECTORY}
mv
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_ascat_outputs/${FILE_PREFIX}_ascat
purity.txt ${PATH_WORKING_DIRECTORY}
```

```
date
module purge
echo "finished ascat"

# CN calling with QDNAseq
# samtools version
module purge; module load bluebear
module load bear-apps/2021b
module load SAMtools/1.15.1-GCC-11.2.0
module load R/4.2.0-foss-2021b
R CMD BATCH ${PATH_QDNA_SEQ_R_FILE}
date
module purge
echo "finished qDNA seq"

date
echo "Finished structural variant calling part1"
echo "################################"

END OF STRUCTURAL VARIANT CALLING PART1
########################################

VCF FORMATTING STEPS FOR CLAIR3 OUTPUTS AND OTHERS
date
echo "Beginning VCF formatting for clair3"

EXPLANATION OF SOFTWARE

# tabix:
# the -p vcf means the input files are vcf files to be indexed so they can
be processed quicker by future programmes,
# file.vcf.tbi is the output extension.
#
# bcftools:
# bcftools concat will concatenate multiple BCF or VCF files into a single
file.
# The difference between bcftools concat and bcftools merge is concat will
concatenate VCF/BCF files from the same set of samples,
# whereas merge will merge VCF/BCF files files from non-overlapping sample
sets
# -a stands for allow overlaps, First coordinate of the next file can
precede last record of the current file,
# this is useful when you want to merge files that might have variants that
are not perfectly aligned.
# -O followed by a lowercase b,u,z or v specifies the output file type,
# Output compressed BCF (b), uncompressed BCF (u), compressed VCF (z),
uncompressed VCF (v). This line is -Oz which means a .vcf.gz output.
# -o: This option allows you to specify the name of the output file that
will be created from the concatenation.
#
### Order of command line:
# tabix -p vcf <input vcf file>
# bcftools concat -a -Oz -o <output file name> <input file 1> <input file
2>
# bcftools sort -Oz -o <output file name> <input file>
#
### Ultimate outputs:
# ${FILE_PREFIX}.clair3.indel.vcf.gz.tbi,
${FILE_PREFIX}.clair3.snv.vcf.gz.tbi,
${FILE_PREFIX}_clair_combinedsomaticcalls.sorted.vcf.gz,
${FILE_PREFIX}_clair_combinedsomaticcalls.sorted.vcf.gz.tbi |
```

```
${FILE_PREFIX}.sniffles2.vcf.tbi | ${FILE_PREFIX}.cutesv.vcf.tbi |
${FILE_PREFIX}.svim.vcf.tbi | ${FILE_PREFIX}.straglr-genotype.vcf.tbi

##############################################################################

module purge; module load bluebear
module load bear-apps/2023a
module load BCFtools/1.20-GCC-12.3.0
#tabix version 1.20
tabix -p vcf ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.clair3.indel.vcf.gz
tabix -p vcf ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.clair3.snv.vcf.gz

#bcftools version 1.20 using htslib version 1.20
bcftools concat -a -Oz -o
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_clair_combinedsomaticcalls.vcf.gz
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.clair3.indel.vcf.gz ${PATH_WORKING
_DIRECTORY}/${FILE_PREFIX}.clair3.snv.vcf.gz
bcftools sort -Oz -o
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_clair_combinedsomaticcalls.sorted.
vcf.gz ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_clair_combinedsomaticcalls.
vcf.gz
date
echo "finished bcftools"

#!!! this line I have removed from the original script as it is
wrong:  mv clairs-to/merged.vcf.gz clairs-to/combinedsomaticcalls.vcf.gz
rm
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_clair_combinedsomaticcalls.vcf.gz
#remove the unsorted version of the clair vcf file

tabix -
p vcf ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_clair_combinedsomaticcalls.s
orted.vcf.gz


# now index the vcf files for the other outputs of the structural variant
software
tabix -p vcf ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.sniffles2.vcf
tabix -p vcf ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.cutesv.vcf
tabix -p vcf ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.svim.vcf
tabix -p vcf ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.straglr-genotype.vcf

date
module purge
echo "Finished VCF formatting for clair3 and other structural
variant vcf outputs"
echo "################################"

END OF VCF FORMATTING STEPS FOR CLAIR3 OUTPUTS

STRUCTURAL VARIANT CALLING PART 2

date
echo "Beginning structural variant calling part 2"

EXPLANATION OF SOFTWARE
# all of these software require you to have
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_clair_combinedsomaticcalls.sorted.
vcf.gz from clair3 as input
# spectre:
```

```
# requires SNPs from clair output and the coverage from
the mosdepth outputs to detect copy number variation
#
# CNVpytor:
#
#
# VCF2MAF:
#
#
# A MAF file identifies, for all samples in a project, the discovered
putative or validated mutations and categorizes those mutations
(polymorphism, deletion, or insertion) as somatic (originating in
the tumor tissue) or germline (originating from the germline).
#
# Order of command line:
# spectre CNVCaller --coverage <path to coverage bed file> --sample-id
<sample ID> --output-dir <output directory> --reference <path to
reference fasta> --metadata <path to MDR metadata>--blacklist <path to
blacklist> --snv <input clair vcf snv>
# cnvpytor ????!!!
# vcf2maf.pl ????!!!

module purge;module load bluebear
# CN calling with Spectre - spectre2 version 0.2.1
${COMMAND_CONDA_ACTIVATE}/spectre2
spectre CNVCaller --coverage
${PATH_WORKING_DIRECTORY}/mosdepth/${FILE_PREFIX}.regions.bed.gz --sample-
id ${FILE_PREFIX} --output-
dir ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_spectre2/ --reference
${PATH_REF_GENOME_FA} --metadata ${PATH_REF_GENOME_SPECMET_MDR} --blacklist
${PATH_REF_GENOME_SPECLIS_BED} --
snv ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_clair_combinedsomaticcalls.sor
ted.vcf.gz
${COMMAND_CONDA_DEACTIVATE}
date
echo "finished spectre"

# CNVpytor version 1.3.1
module purge; module load bluebear
${COMMAND_CONDA_ACTIVATE}/cnvpytor
# I don't know if these are in the correct order, the order on github is
different
# maybe you could even put all of this onto the onle line of code instead?
cnvpytor -root ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.pytor -
snp ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_clair_combinedsomaticcalls.sor
ted.vcf.gz
cnvpytor -root ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.pytor -
rd ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.aligned.sorted.bam
cnvpytor -root ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.pytor -mask_snps
cnvpytor -root ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.pytor -baf 10000
cnvpytor -root ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.pytor -his 10000
cnvpytor -root ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.pytor -call
combined 10000 > ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_CNVPytor.tsv
${COMMAND_CONDA_DEACTIVATE}
date
echo "finished cnvpytor"


# Run VCF2MAF
# original script makes a sorted uncompressed bed file as input, but
I don't know why that would have to happened when gzip -d exists?
```

```
# also could check if vcf2maf has a feature to take .gz
    # vcf2maf version 1.6.18 and perl version 5.30.0
# the ${EBROOTVEP} stands for: /rds/bear-apps/2019b/EL8-
    cas/software/VEP/99.2-foss-2019b-Perl-5.30.0
                                cp
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_clair_combinedsomaticcalls.sorted.
vcf.gz ${PATH_WORKING_DIRECTORY}/FOR_VCF2MAF_${FILE_PREFIX}_clair_combineds
                    omaticcalls.sorted.vcf.gz
                                bgzip -d
${PATH_WORKING_DIRECTORY}/FOR_VCF2MAF_${FILE_PREFIX}_clair_combinedsomaticc
    alls.sorted.vcf.gz    # makes the uncompressed version without .gz
                    #module purge;module load bluebear
          #module load vcf2maf/1.6.18-foss-2019b-Perl-5.30.0
# !!! I have no idea what the vep-path or the filter-vcf paths are supposed
                                to be
                            #vcf2maf.pl --input-
vcf ${PATH_WORKING_DIRECTORY}/FOR_VCF2MAF_${FILE_PREFIX}_clair_combinedsoma
                        ticcalls.sorted.vcf --output-
  maf ${PATH_WORKING_DIRECTORY}/FOR_VCF2MAF_${FILE_PREFIX}.maf --vep-path
 ${EBROOTVEP} --vep-data ${PATH_VEP_DATA} -filter-vcf ${PATH_VEP_DATA_EXAC}
--ncbi-build GRCh38 --cache-version 97 --ref-fasta ${PATH_REF_GENOME_FA} --
 vep-forks 12 --tumor-id ${FILE_PREFIX} --vcf-tumor-id SAMPLE --vcf-normal
                                SAMPLE

                                date
                            #echo "finished vcf2maf"

                                date
             echo "Finished structural variant calling part 2"
                  echo "################################"
                  END OF STRUCTURAL VARIANT CALLING PART 2

                            CSPR AND REPORTING
                                date
              echo "Beginning CSPR and reporting results"

                        EXPLANATION OF SOFTWARE
              # all of these software require you to have
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_clair_combinedsomaticcalls.sorted.
                        vcf.gz from clair3 as input
    # Tabix, bcftools, sed and bgzip are all just for file conversions
                                # tabix:
# -f: Forces an index to be made for the input, overwriting existing index
                                files.
                # -B: input file is a "bgzip" compressed file
# -h: include the header lines from the input file in the indexed output.
  # the entire of the first command means that the clair vcf file will be
   filtered to only include the regions in the truseq_hg38 file and then
                        create a new vcf file output
          # the -p vcf line means the input vcf will be indexed
                                #
                              # bcftools:
# bcftools view -i 'FILTER="NonSomatic"' means to extract the "NonSomatic"
                values and then put them into a new vcf file
                                #
                                # sed:
      # sed -i means that the file you give will be overwritten, for
example sed -i 's/NonSomatic/PASS/g' germlinevariantsonly.vcf will take all
 the instances ("g"=global) of "NonSomatic" and substitute ("s") these for
                                "PASS"
      # The above sed is standard usage format and is case sensitive
```

```
#
# Awk
# It reads the file ascatploidy.txt, checks each line to see if the second
field starts with a digit, and if it does, it prints that second field. The
output of this is then piped to cut, which takes the first four characters
of each printed second field. The final result is the output of this entire
process, which will be a list of the first four characters of the second
fields from lines where those fields start with a digit.

##############################################################################

# CSPR and PCGR reporting
# pcgr/cpsr version 2.1.2
#
module purge; module load bluebear
module load bear-apps/2023a
module load BCFtools/1.20-GCC-12.3.0

tabix -
fBh ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_clair_combinedsomaticcalls.sor
ted.vcf.gz ${PATH_REF_GENOME_TRUSEQ_HG38_BED} >
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_exome_filtered.vcf
bgzip ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_exome_filtered.vcf
tabix -
p vcf ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_exome_filtered.vcf.gz
bcftools view -i 'FILTER="NonSomatic"' -Ov -o
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_germlinevariantsonly.vcf ${PATH_WO
RKING_DIRECTORY}/${FILE_PREFIX}_exome_filtered.vcf.gz
bcftools view -i 'FILTER="PASS"' -Oz -o
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_somaticvariantsonly.vcf.gz ${PATH_
WORKING_DIRECTORY}/${FILE_PREFIX}_exome_filtered.vcf.gz
tabix -
p vcf ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_somaticvariantsonly.vcf.gz
# gunzip ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_germlinevariantsonly.vcf.
gz
sed -i 's/NonSomatic/PASS/g'
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_germlinevariantsonly.vcf
bgzip ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_germlinevariantsonly.vcf
tabix -
p vcf ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_germlinevariantsonly.vcf.gz
Tumour_ploidy=$(awk '{ if ($2 ~ /^[0-9]/) print $2 }'
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_ascatploidy.txt | cut -c1-4)
export Tumour_ploidy
Tumour_purity=$(awk '{ if ($2 ~ /^[0-9]/) print $2 }'
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_ascatpurity.txt | cut -c1-4)
export Tumour_purity
cat ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_ascat.segments.txt | awk
'{print $2,$3,$4,$5,$6}' | tr ' ' '\t' | sed 's/chr/Chromosome/g'
| sed 's/startpos/Start/g' | sed 's/endpos/End/g' >
${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.segments2.txt

module purge
${PCGR_ACTIVATION}
cpsr --
input_vcf ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_germlinevariantsonly.vcf
.gz --output_dir ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_cpsr/ --
genome_assembly grch38 --sample_id ${FILE_PREFIX} --panel_id 0 --
vep_dir ${VEP_DIR} --refdata_dir ${REF_DATA_PCGR} --secondary_findings --
classify_all --force_overwrite
echo "cpsr done"
```

```
    pcgr --output_dir ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_pcgr/ --
            sample_id ${FILE_PREFIX} --genome_assembly grch38 --
input_vcf ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_somaticvariantsonly.vcf.
 gz --assay WES --vep_n_forks 4 --vep_no_intergenic --vep_dir ${VEP_DIR} --
                        refdata_dir ${REF_DATA_PCGR} --
input_cpsr ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_cpsr/${FILE_PREFIX}.cps
                     r.grch38.classification.tsv.gz --
input_cpsr_yaml ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}_cpsr/${FILE_PREFIX
}.cpsr.grch38.conf.yaml --estimate_signatures --all_reference_signatures --
        include_artefact_signatures --estimate_tmb --estimate_msi --
    input_cna ${PATH_WORKING_DIRECTORY}/${FILE_PREFIX}.segments2.txt --
             force_overwrite --tumor_ploidy ${Tumour_ploidy} --
                       tumor_purity ${Tumour_purity}
                        ${COMMAND_CONDA_DEACTIVATE}

                             ${CONDA_DEACTIVATE}
                                      #
                                      #
                                    date
                  echo "Finished CSPR and reporting results"
                    echo "################################"
                                    date
                        echo "Finished all the script!"
```

Appendix 5: Single-cell – cell annotation pipeline

```
library(Seurat)
library(reticulate)
library(anndata)
library(ggplot2)

###### 1) load in the datasets
# The query and atlas datasets have been put through threshold filtering, normalised, FindVariables and scaled as in previous steps with Seurat

atlas1 <- readRDS("Filtered_normalised_findvariables_scaled-human_limb_embryo_atlas.RDS")
atlas2 <- readRDS("Filtered_normalised_findvariables_scaled-ossification_atlas.RDS")
my_query <- readRDS("High_depth_sample_010156-seurat_object_filtered_normalised_findvariables_scaled.RDS")

####### 2) all the data filtering and processing has been done already, now find the anchours

anchors_atlas1 <- FindTransferAnchors(reference = atlas1, query = my_query)
anchors_atlas2 <- FindTransferAnchors(reference = atlas2, query = my_query)

####### 3) transfer labels
# in the human limb atlas the cell annotations are listed under the "cell_type" column of its metadata
# in the ossification atlas the cell annotations are listed under the "fineanno" column of its metadata

predictions1 <- TransferData(anchorset = anchors_atlas1, refdata = atlas1$cell_type)
my_query <- AddMetaData(object = my_query, metadata = predictions1$predicted.id, col.name = "predictions_atlas1")
my_query <- AddMetaData(object = my_query, metadata = predictions1$prediction.score.max, col.name = "predictions_atlas1_score")

predictions2 <- TransferData(anchorset = anchors_atlas2, refdata = atlas2$fineanno)
my_query <- AddMetaData(object = my_query, metadata = predictions2$predicted.id, col.name = "predictions_atlas2")
my_query <- AddMetaData(object = my_query, metadata = predictions2$prediction.score.max, col.name = "predictions_atlas2_score")

saveRDS(my_query, "High_depth_sample_010156-WITH_PREDICTED_CELL_TYPES.RDS")

########## Making graphs of the results and saving as csv
seurat_object <- readRDS("High_depth_sample_010156-WITH_PREDICTED_CELL_TYPES.RDS")

seurat_object$best_prediction <- ifelse(seurat_object$predictions_atlas1_score > seurat_object$predictions_atlas2_score, seurat_object$predictions_atlas1, seurat_object$predictions_atlas2)  # the atlas annotation with the highest score per cell gets put into a new column
seurat_object$best_prediction_score <- ifelse(seurat_object$predictions_atlas1_score > seurat_object$predictions_atlas2_score,
```

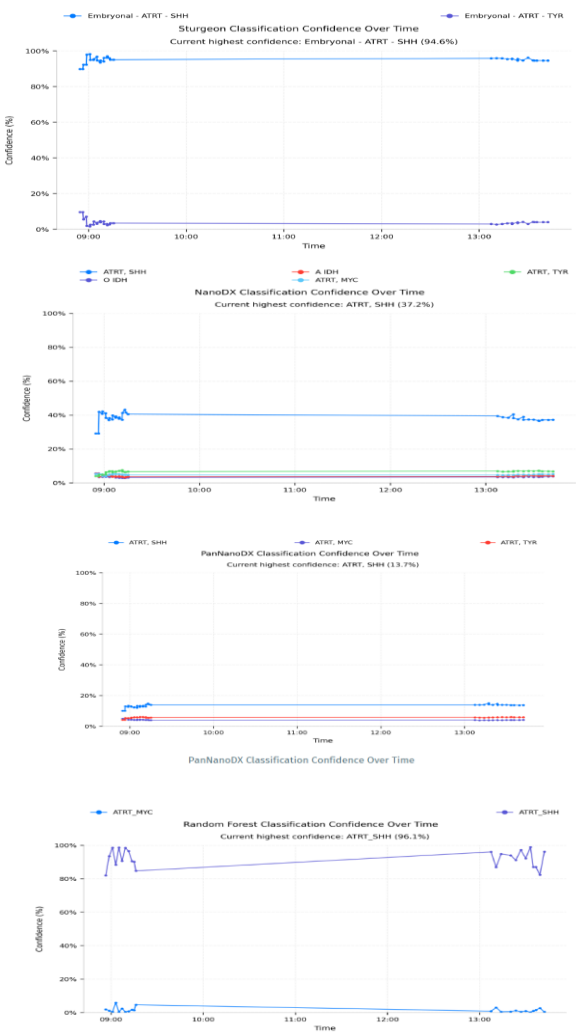
```
                    seurat_object$predictions_atlas1_score,
seurat_object$predictions_atlas2_score)  # the highest atlas score gets put
                                    into a column

# so that you can actually plot the scatter with the new coloured points of
            the different cell annotations, here is what you do:
          # do the PCA and UMAP stuff to generate the coordinates

          seurat_object <- RunPCA(object = seurat_object, features
                = VariableFeatures(object = seurat_object))
 seurat_object <- FindNeighbors(object = seurat_object, dims = 1:9) # the
    dims needs to be set for 1 and up to the result of the elbow curve

                        #### GRAPH CLUSTERING WITH UMAP
 seurat_object <- FindClusters(object = seurat_object, resolution = c(0.1,
   0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1))  # this will find all the
resolution types and you can change the res number in the dimplot to any of
                        these to generate the right file
                #plots the resolutions over PCA of PC1 vs PC2
      ########PLOTTING THE UMAP AND FINAL DETERMINATION OF RESOLUTION
# the number of dimensions needs to be the same number as for the PCA elbow
       curve (NOT the number of clusters defined by the resolution)
       seurat_object <- RunUMAP(object = seurat_object, dims = 1:9)
             # plot all of your resolutions on top of the UMAP
DimPlot(seurat_object, group.by = "predictions_atlas1", pt.size = 2, cols =
                                    "polychrome")
DimPlot(seurat_object, group.by = "predictions_atlas2", pt.size = 2, cols =
                                    "polychrome")
  DimPlot(seurat_object, group.by = "best_prediction", pt.size = 2, cols =
                                    "polychrome")
          meta_data_df <- as.data.frame(seurat_object@meta.data)
            write.csv(meta_data_df, "High_depth_sample_010156-
                        WITH_PREDICTED_CELL_TYPES.csv")
 # making a subset so that you can see what the dimplot looks like but with
          only the cells that you have confidence in annotating
  seurat_object <- subset(seurat_object, subset = best_prediction_score >
                                        0.8)
  DimPlot(seurat_object, group.by = "best_prediction", pt.size = 2, cols =
                                    "polychrome")
```

Appendix 6: Methylation classifier output

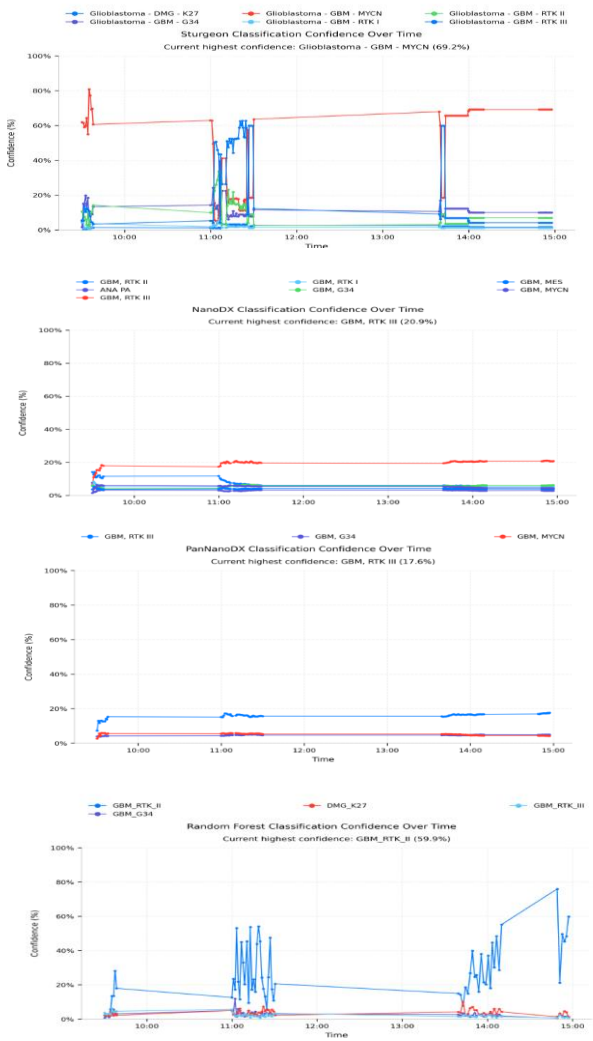
Figure A. 1 Methylation classifiers outputs generated from ROBIN pipeline for all cases. *The plots show the classification confidence over time for four classifiers: Sturgeon, NanoDX, PanNanoDx, and Random Forest.*

NP006-2017

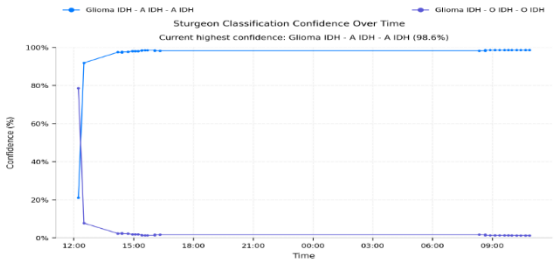
NP026-2020

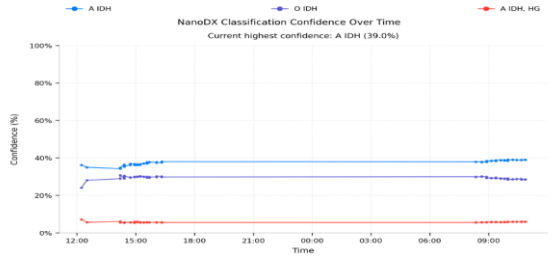


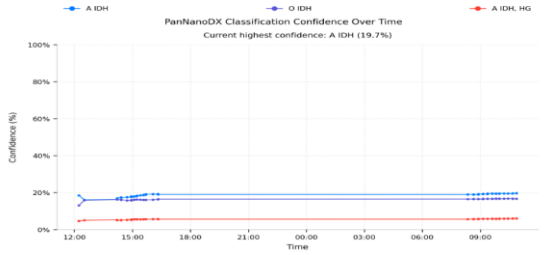
NP039-2020

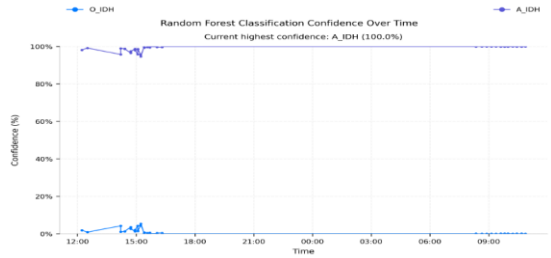


NP018-2018

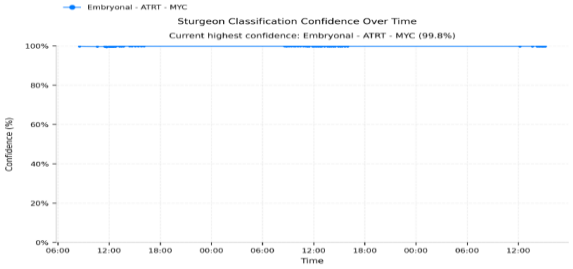


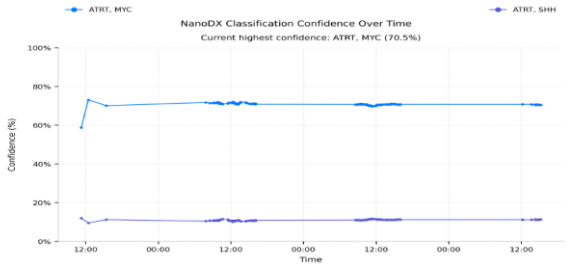


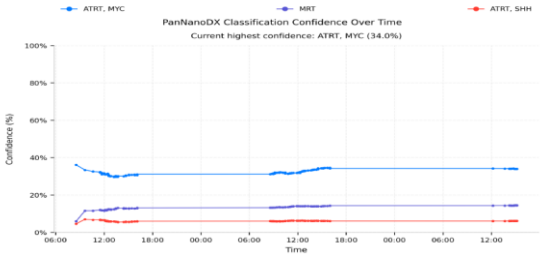


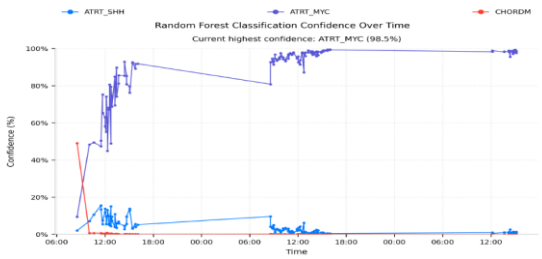


NP0051-2017

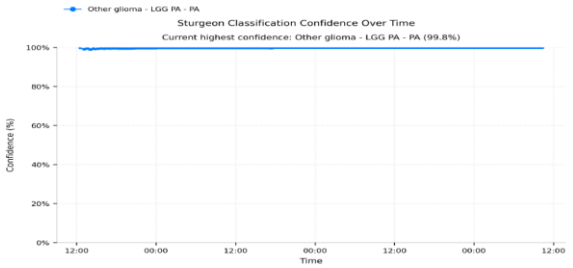


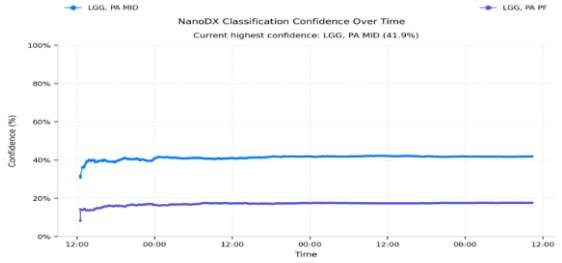


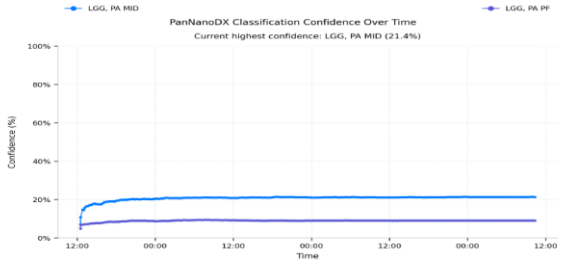


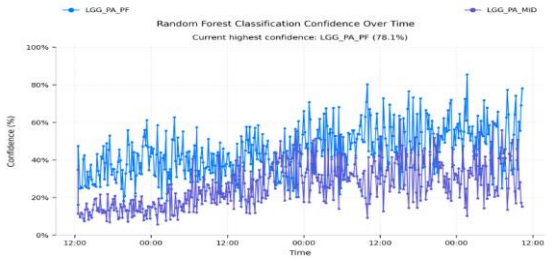


NP063-2022

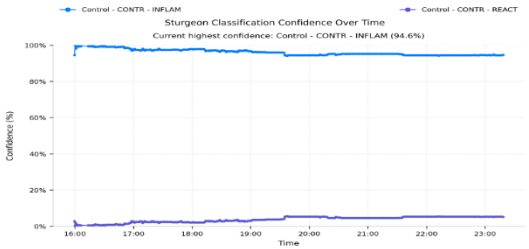


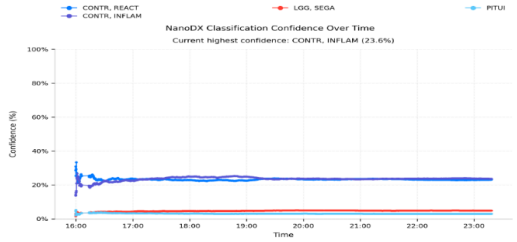


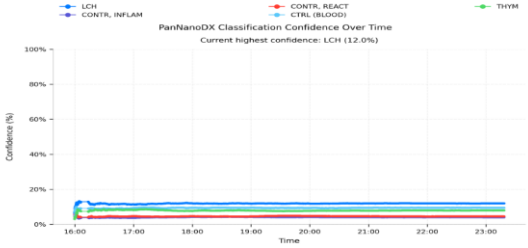


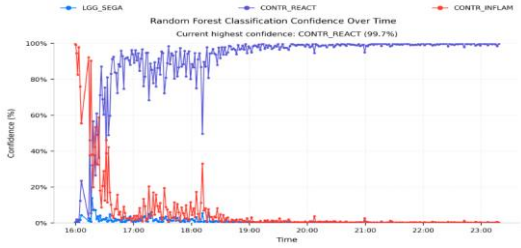


SH214-2021

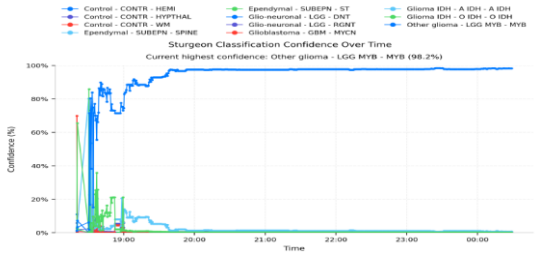


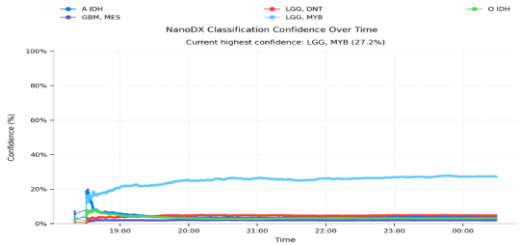


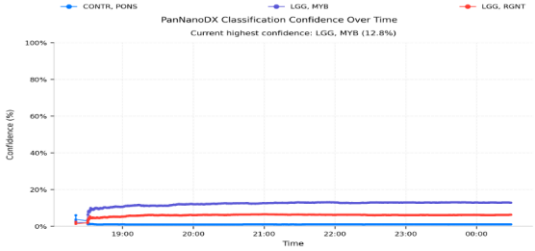


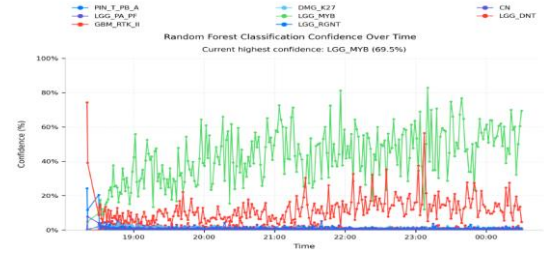


SH451-2023

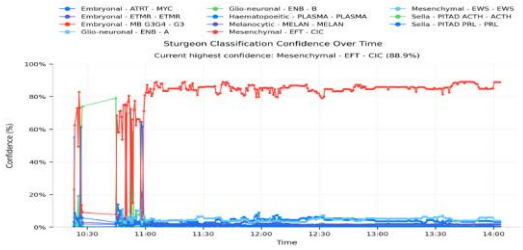


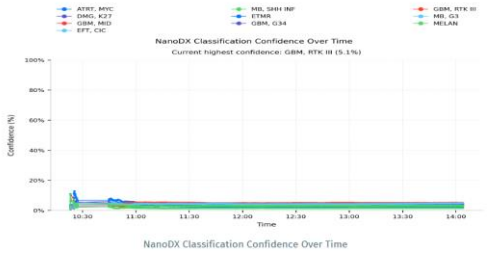


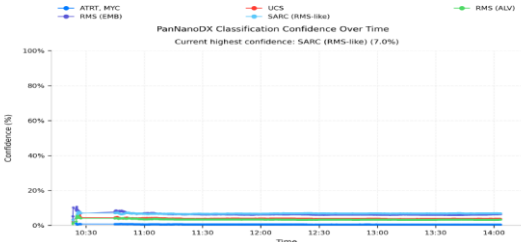


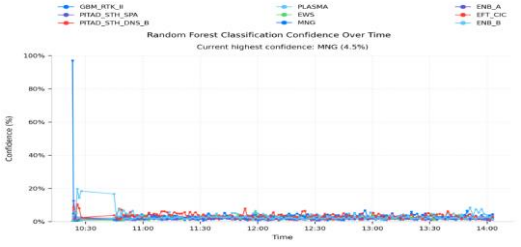


SH576-2023

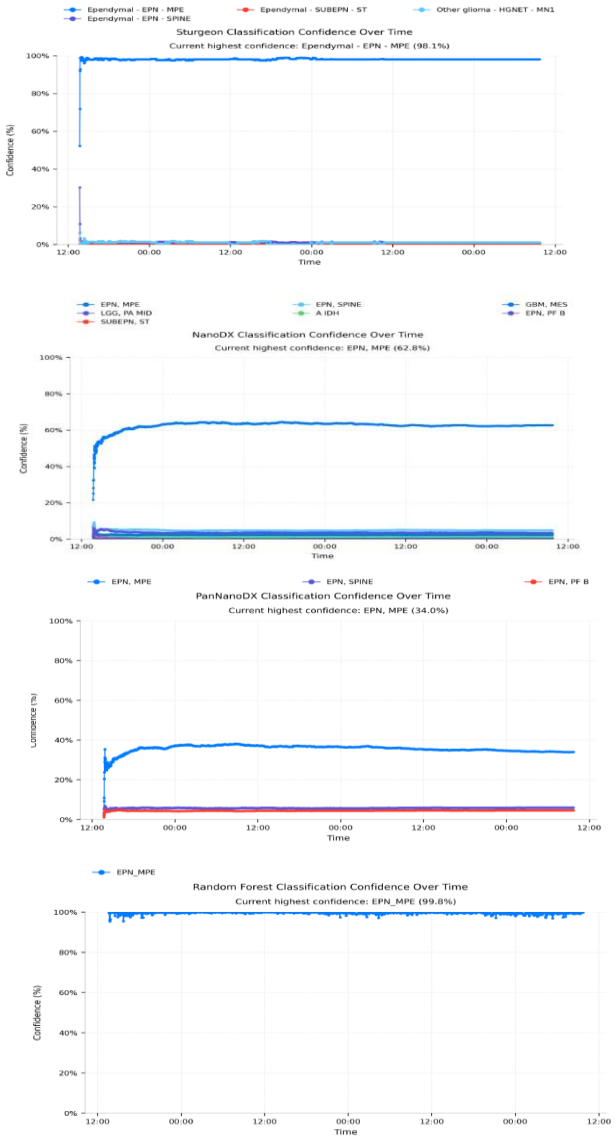




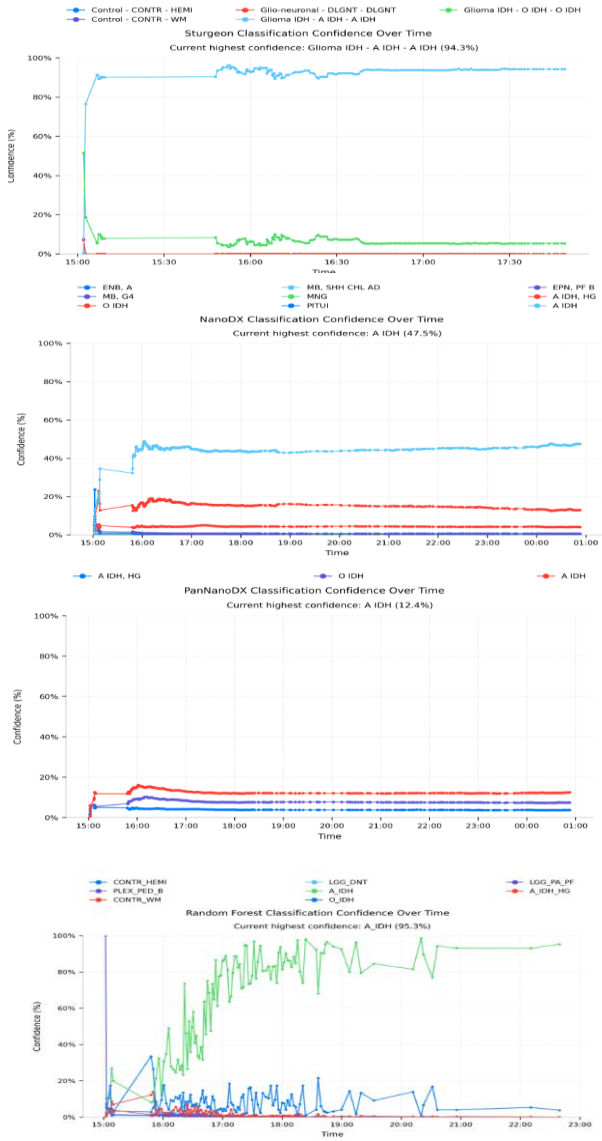




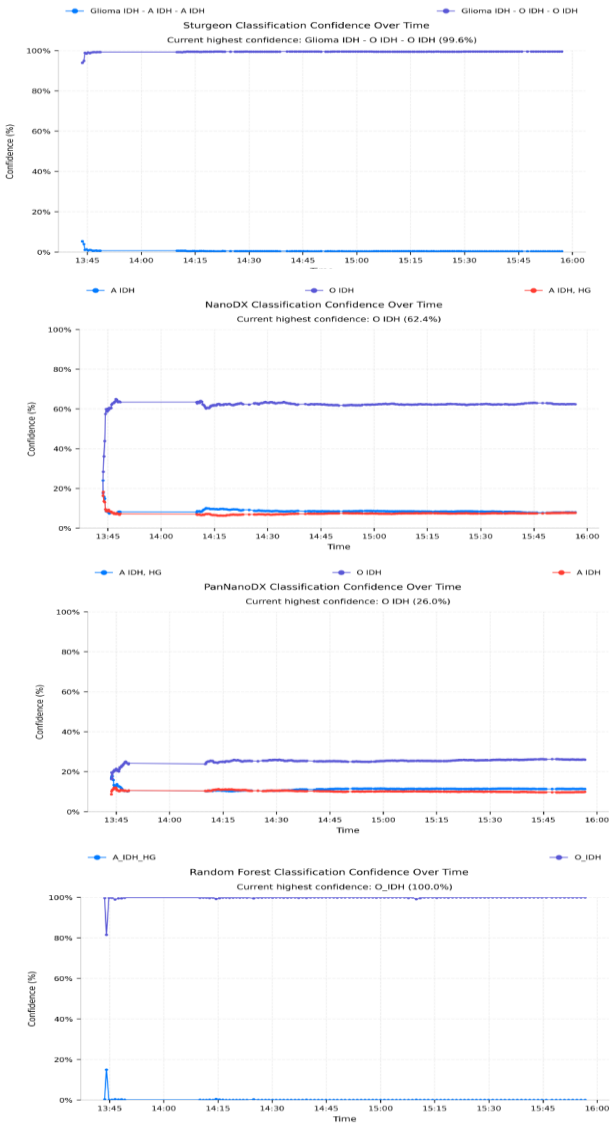
SH894-2025



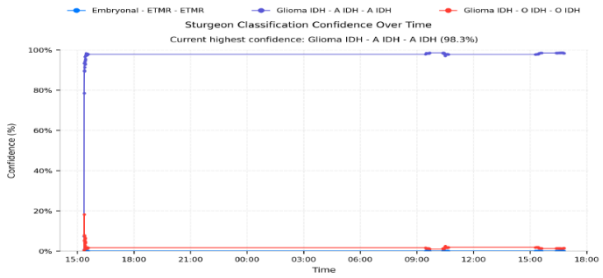
SH912-2025

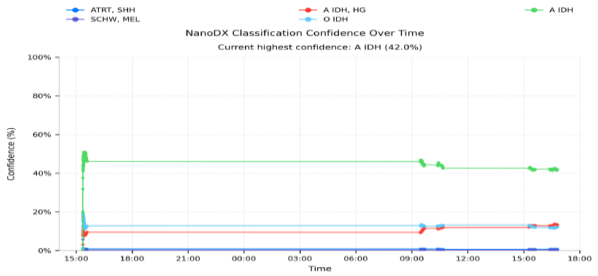


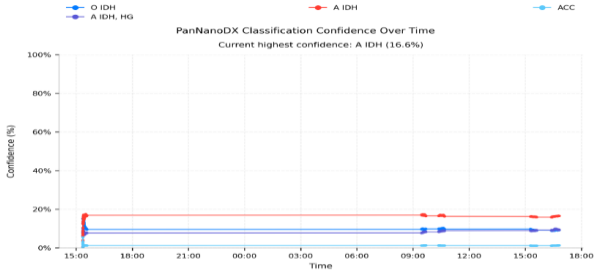
SH1343-2020

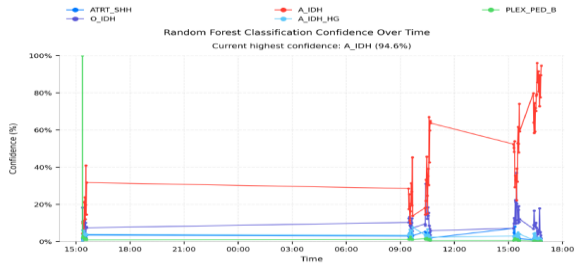


SH1049-2025


Embryonal - ETMR - ETMR
Glioma IDH - A IDH - A IDH
Glioma IDH - O IDH - O IDH
Sturgeon Classification Confidence Over Time
Current highest confidence: Glioma IDH - A IDH - A IDH (98.3%)
Confidence (%)
Time


ATRT, SHH
SCHW, MEL
A IDH, HG
O IDH
A IDH
NanoDX Classification Confidence Over Time
Current highest confidence: A IDH (42.0%)
Confidence (%)
Time

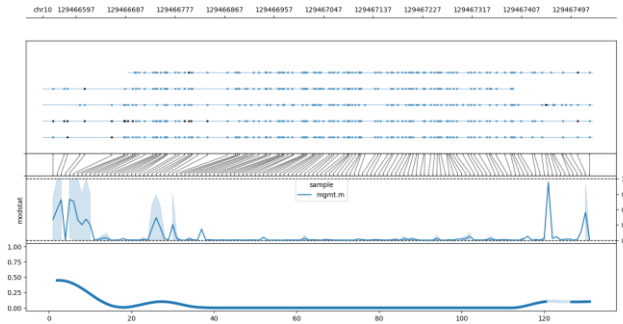

O IDH
A IDH, HG
A IDH
ACC
PanNanoDX Classification Confidence Over Time
Current highest confidence: A IDH (16.6%)
Confidence (%)
Time


ATRT_SHH
O_IDH
A_IDH
A_IDH_HG
PLEX_PED_B
Random Forest Classification Confidence Over Time
Current highest confidence: A_IDH (94.6%)
Confidence (%)
Time

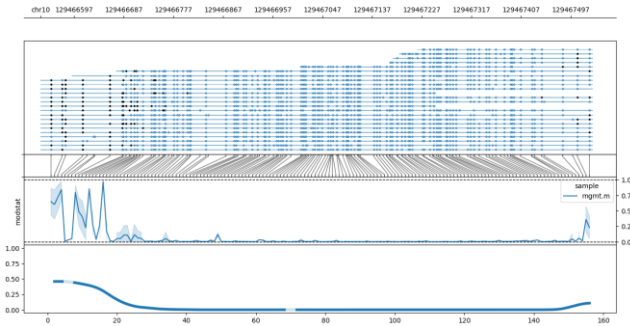
Appendix 7: ROBIN MGMT promotor methylation plot showing methylation levels across CpG sites

***Figure A. 2* MGMT promoter methylation status**. *MGMT promoter is located on Chromosome 10, and each dot represents a CpG site within that region. The blue lines correspond to a single read. The black dots indicate CpG methylated sites and the blue dots are unmethylated sites.*

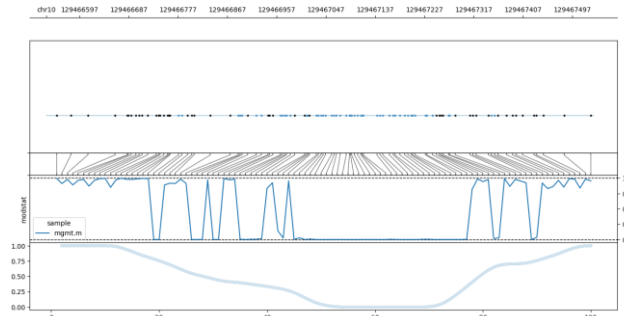
NP006-2017



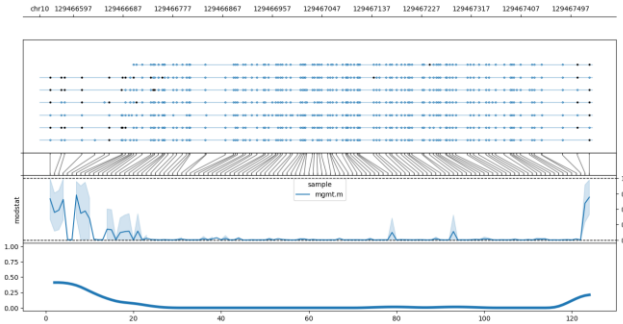
NP026-2019



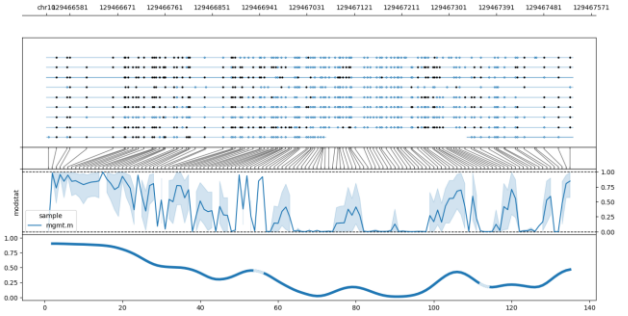
NP039-2020



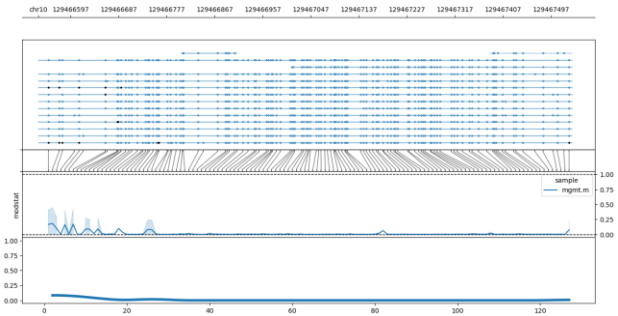
NP063-2022



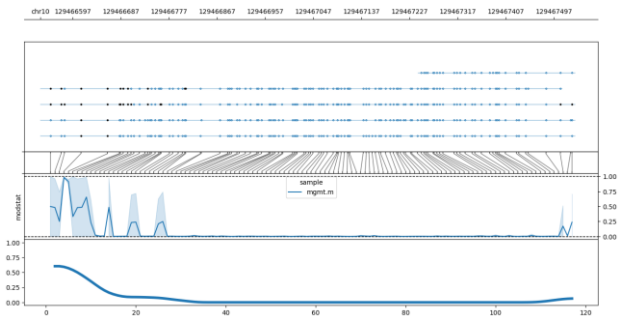
NP018-2018



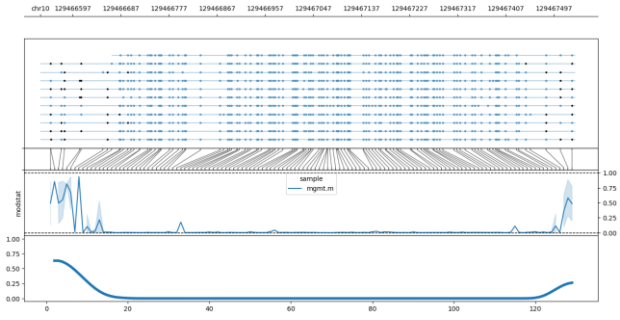
NP051-2017



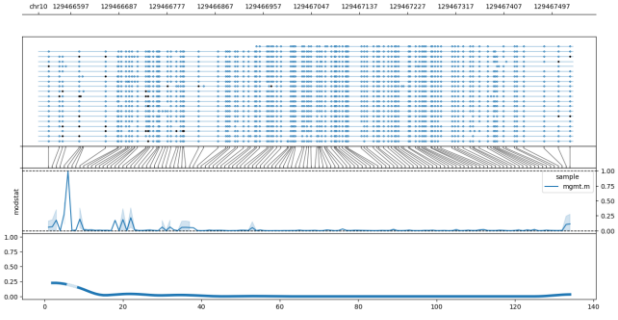
SH214-2021



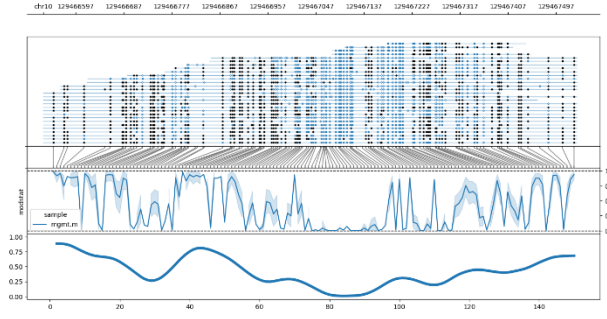
SH451-2023



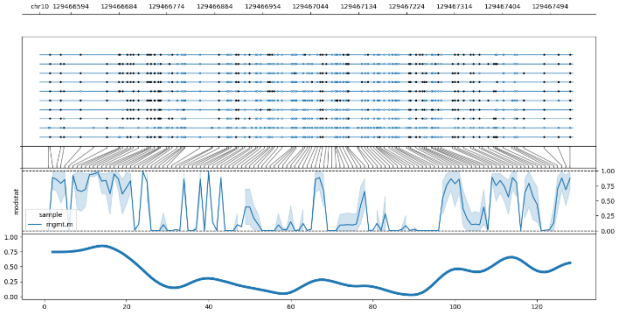
SH576-2023



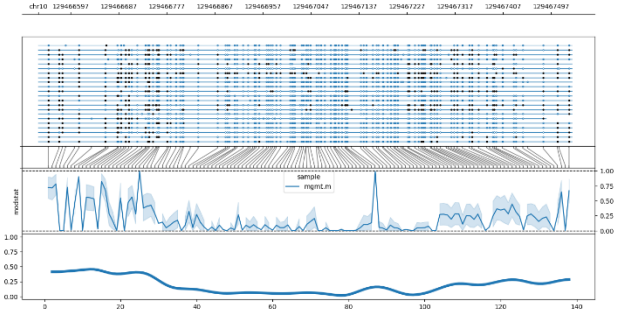
SH1343-2020



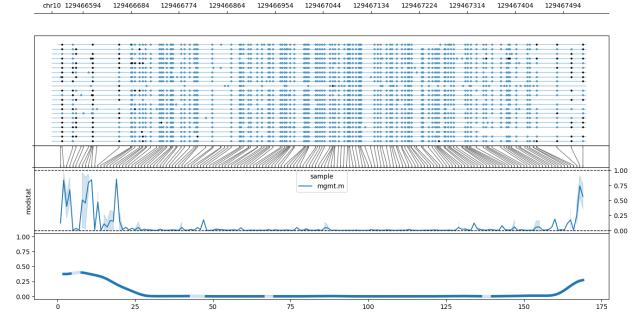
SH894-2025



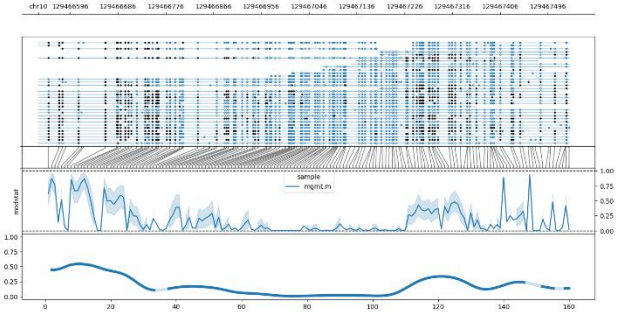
SH912-2025



SH946-2025

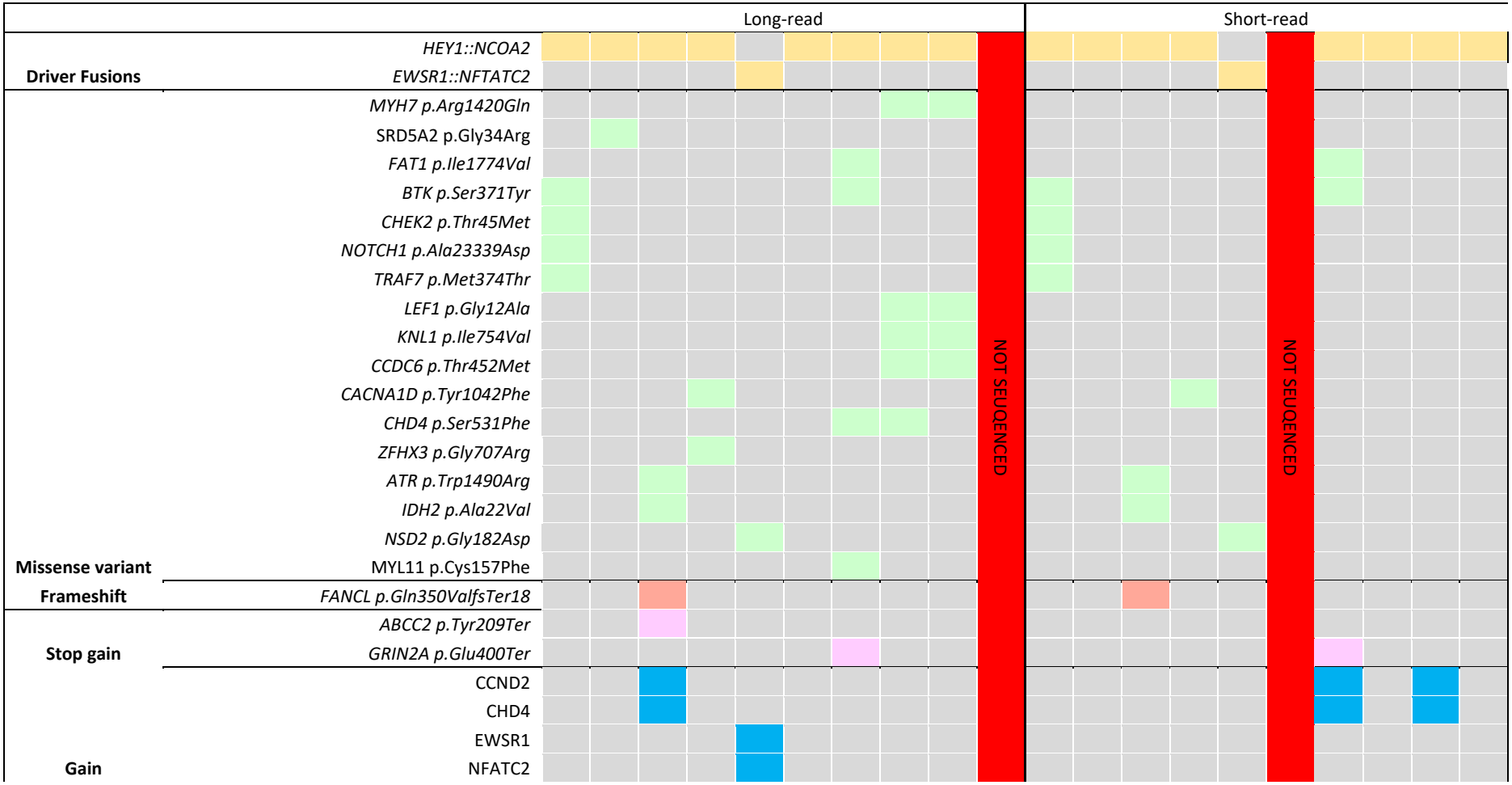


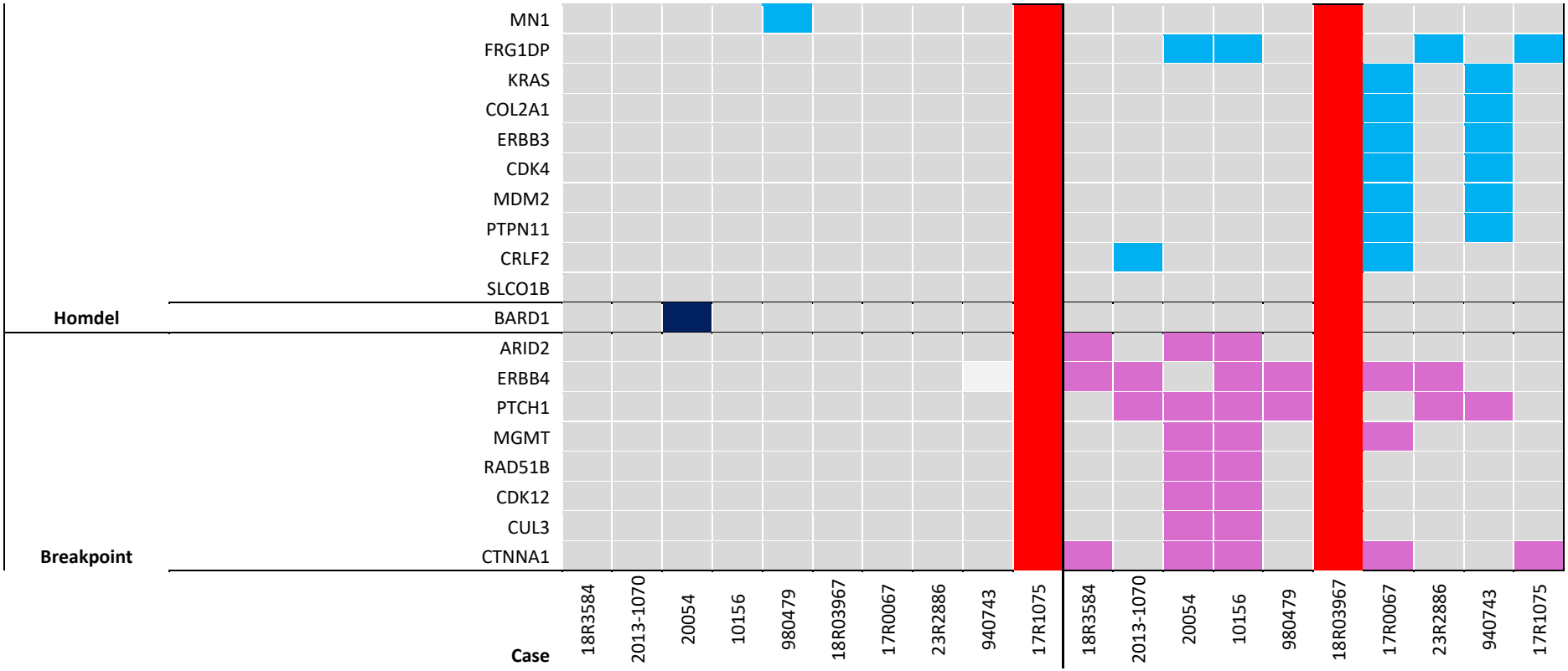
SH1049-2025



Appendix 8: Full table of variants detect in MCS cases with long-read and short-read WGS sequencing

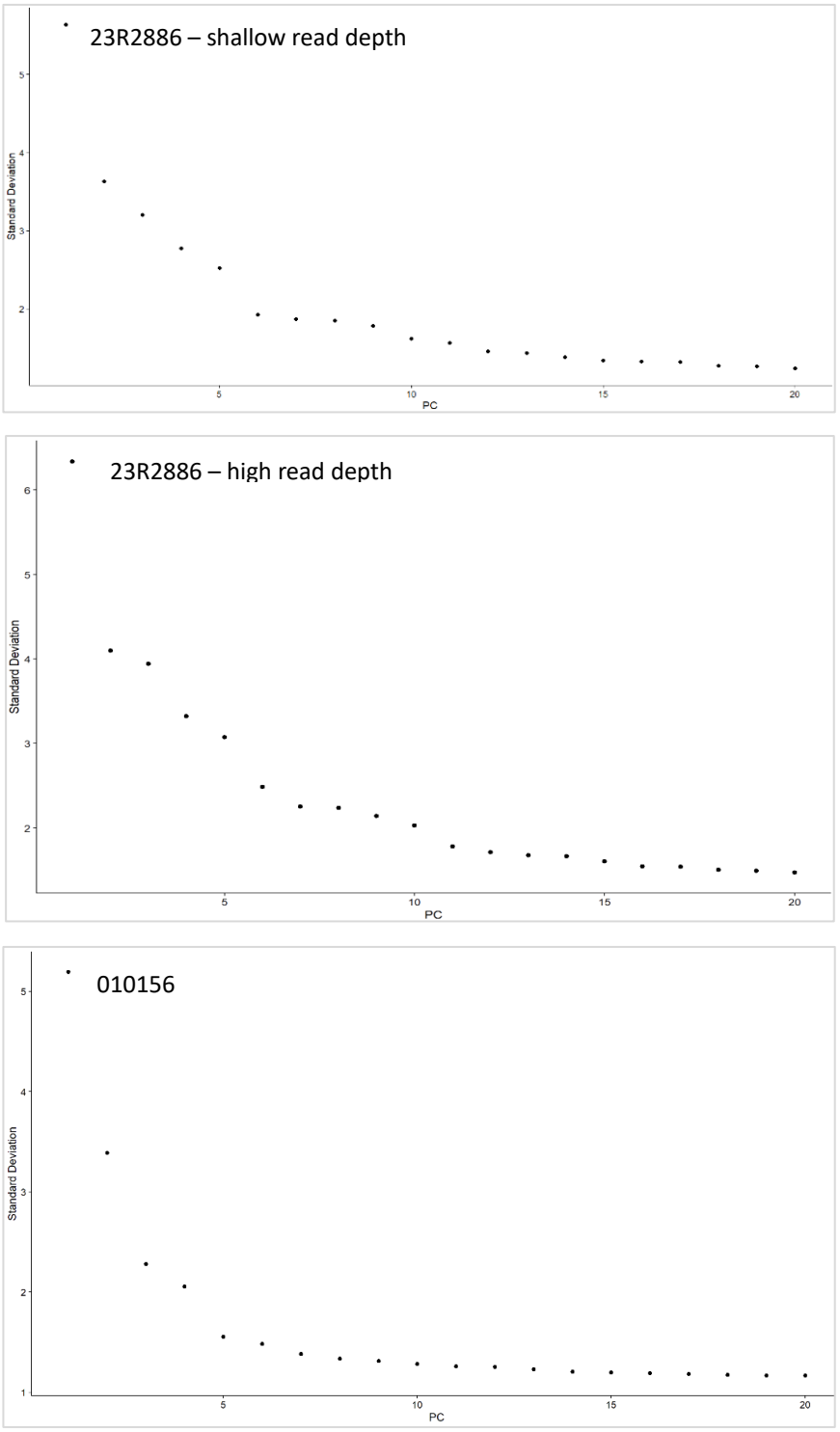
Figure A. 3 Full list of somatic variants detected in MCS cases by long-read and short-read sequencing.




MN1
FRG1DP
KRAS
COL2A1
ERBB3
CDK4
MDM2
PTPN11
CRLF2
SLCO1B
Homdel
BARD1
ARID2
ERBB4
PTCH1
MGMT
RAD51B
CDK12
CUL3
Breakpoint
CTNNA1
Case
18R3584
2013-1070
20054
10156
980479
18R03967
17R0067
23R2886
940743
17R1075
18R3584
2013-1070
20054
10156
980479
18R03967
17R0067
23R2886
940743
17R1075

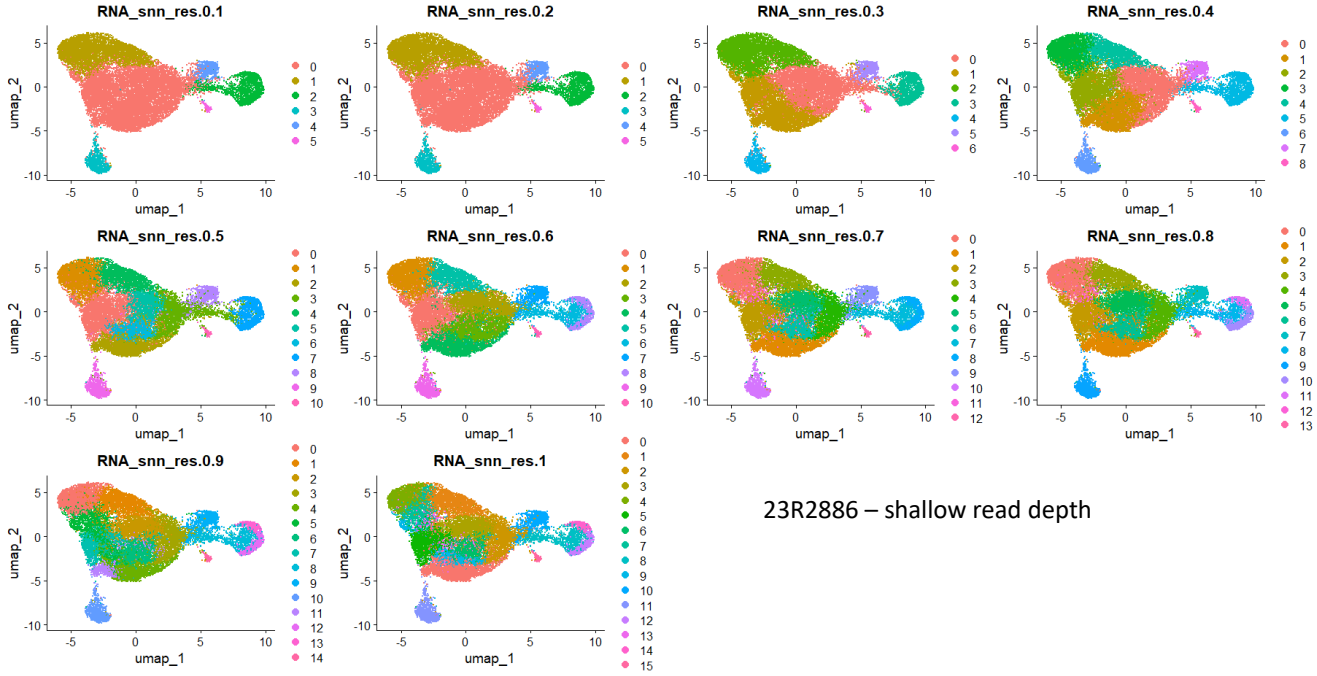
Appendix 9: Elbow plots used to identify number of principle components

Figure A. 4 Elbow plots used to identify number of principle components from Single-cell data. *The plot shows the standard deviation of each principle component and the bend in the graph identifies the threshold for variation.*

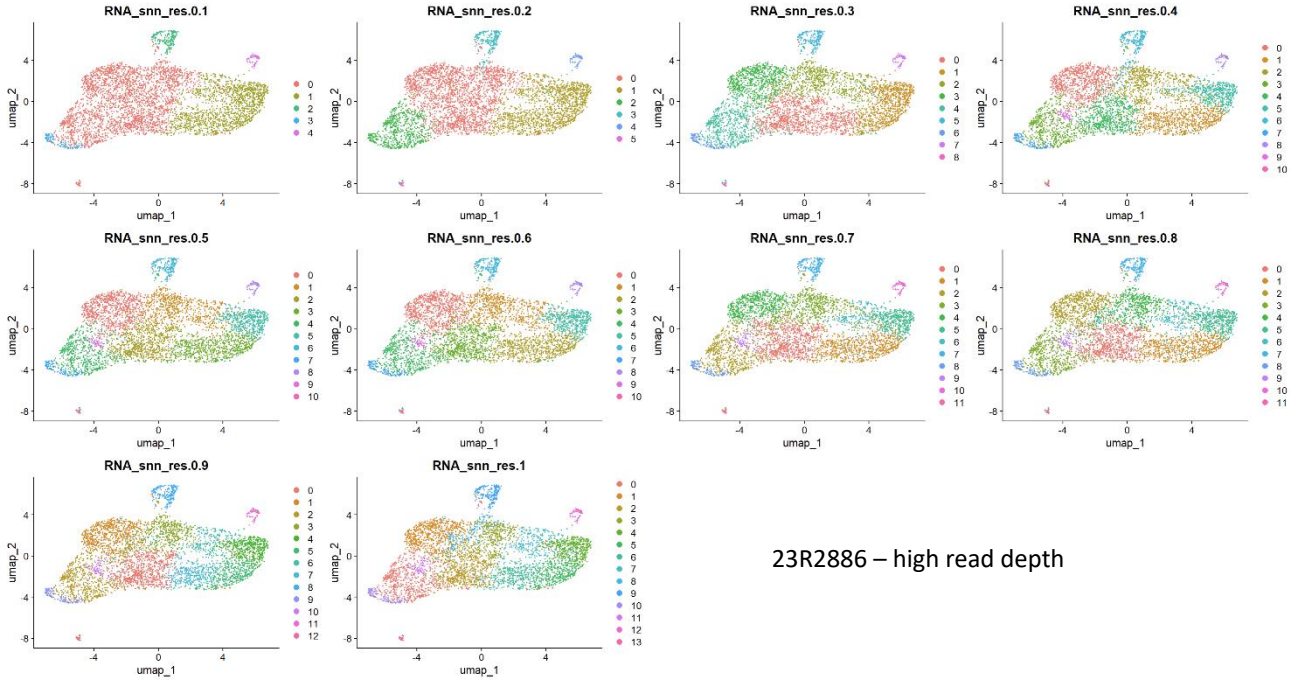


Appendix 10: UMAPs at resolution 0.1 – 1 after thresholding

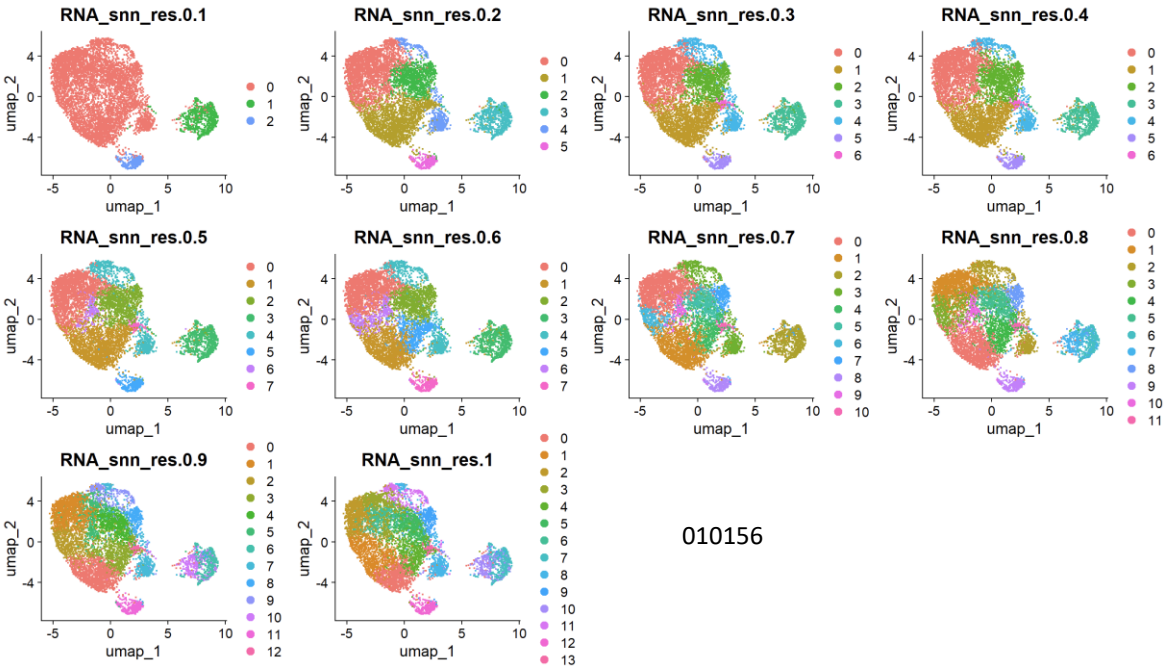
Figure A. 5 Visualisation of single-cell RNA-seq data clustered at resolution 0.1 - 1. *Each point represents an individual cell, positioned by UMAP dimensions 1 and 2 based on transcriptomic similarity. Cells are clustered which demonstrates relationships among major cell populations.*



23R2886 – shallow read depth



23R2886 – high read depth


RNA_snn_res.0.1
RNA_snn_res.0.2
RNA_snn_res.0.3
RNA_snn_res.0.4
RNA_snn_res.0.5
RNA_snn_res.0.6
RNA_snn_res.0.7
RNA_snn_res.0.8
RNA_snn_res.0.9
RNA_snn_res.1
umap_1
umap_2

010156