Inaugural dissertation

for

obtaining the doctoral degree

of the

Combined Faculty of Mathematics, Engineering and Natural Sciences of the

Ruprecht - Karls - University

Heidelberg

Presented by

MRes, Areeba Jamilkhan Patel

Born in: Pune, India

Oral examination: 10th December 2024

Towards accessible molecular diagnostics for central nervous system tumours

Referees: Prof. Dr. Benedikt Brors

Prof. Dr. Dr. med. Felix Sahm

Abstract

The 2021 WHO classification represents a significant shift in Central Nervous System (CNS) tumour diagnostics, emphasising the integration of molecular alterations alongside traditional histopathology. Among the advancements in molecular diagnostics, methylation-based classification using the Heidelberg Molecular Neuropathology (MNP) classifier (molecularneuropathology.org) has become an essential diagnostic tool. Conventional molecular testing often involves multiple assays such as DNA/RNA sequencing, methylation arrays, immunohistochemistry among others, which are resource-intensive and limited to high-throughput settings due to their complexity, costs, and lengthy turnaround times.

In this work, I introduce two tools aimed at improving the accessibility and affordability of CNS tumour molecular diagnostics: Rapid-CNS^2 and MNP-Flex. Rapid-CNS^2 is a nanopore sequencing workflow that employs adaptive sampling to efficiently detect mutations, copy number alterations, gene fusions, target gene methylation, and perform methylation classification, all in a single test. This system is flexible, allowing immediate testing on individual samples and customisable targets via a simple text file. I formulated and subsequently validated the pipeline using 252 samples, including archival and diagnostic frozen sections. I developed ad-hoc models for methylation classification and *MGMT* promoter methylation detection. I employed publicly available state-of-the-art tools for pre-processing, variant calling and annotation, and devised computational acceleration strategies. Additionally, I demonstrate the potential of the pipeline to report results in an intraoperative time-frame with 18 samples from two independent centres. Thus, Rapid-CNS^2 offers real-time methylation classification and DNA copy-number reporting within a 30-minute intraoperative window, followed by comprehensive molecular profiling within 24h, covering the entire spectrum of molecular alterations relevant for diagnosis and targeted therapies for CNS tumour subtypes- drastically reducing the weeks-long turnaround required by conventional methods.

To further enhance accessibility of the MNP classifier, I developed MNP-Flex, a platform-independent version of the MNP classifier, covering 184 CNS tumour classes. I validated MNP-Flex on a global cohort of over 78,000 samples, including both frozen and formalin-fixed

paraffin-embedded (FFPE) samples processed using five different methylation profiling technologies. With clinically relevant thresholds, MNP-Flex achieved accuracies of 99.6% for methylation families and 99.2% for methylation classes.

Together, Rapid-CNS^2 and MNP-Flex offer a comprehensive workflow for CNS tumour diagnostics. Rapid-CNS^2 provides real-time, intraoperative reporting of broad methylation classification and copy number variations to guide surgical strategy, while the complete molecular profile and fine-grained methylation classification with MNP-Flex is available the next day, informing clinical care and therapeutic decisions. The workflow is cost-effective, uses compact equipment, and employs straightforward laboratory and bioinformatics tools. Rapid-CNS^2 is available on GitHub, and MNP-Flex can be accessed via a research-use web service at https://mnp-flex.org. This integrated approach aims to streamline CNS tumour molecular diagnostics, broadening global access to precise, molecularly-informed classification and ultimately improving patient outcomes.

Zusammenfassung

Die WHO-Klassifikation von 2021 markiert einen tiefgreifenden Wandel in der Diagnostik von Tumoren des zentralen Nervensystems (ZNS). Sie legt den Fokus auf die Integration molekularer Veränderungen als Ergänzung zur traditionellen Histopathologie. Besonders hervorzuheben ist der Fortschritt in der Molekulardiagnostik durch die methylierungsbasierte Klassifikation mittels des Heidelberg Molecular Neuropathology (MNP)-Klassifikators (molecularneuropathology.org), der mittlerweile als unverzichtbares Diagnosewerkzeug gilt. Herkömmliche molekulare Analysen setzen häufig mehrere komplexe Methoden wie DNA-/RNA-Sequenzierung, Methylierungsarrays und Immunhistochemie ein. Diese Verfahren sind ressourcenintensiv und aufgrund ihrer Komplexität, Kosten und langen Bearbeitungszeiten auf hochspezialisierte Einrichtungen beschränkt.

In dieser Arbeit präsentiere ich zwei Tools, die darauf abzielen, die Zugänglichkeit und Erschwinglichkeit der molekularen Diagnostik von ZNS-Tumoren zu verbessern: Rapid-CNS^2 und MNP-Flex. Rapid-CNS^2 ist ein Nanoporen-Sequenzierungsworkflow, der mithilfe adaptiven Samplings effizient Mutationen, Kopienzahlveränderungen, Genfusionen und Zielgenmethylierungen erfasst und eine umfassende Methylierungsklassifikation ermöglicht – alles in einem einzigen Workflow. Das System ist flexibel und erlaubt sofortige Analysen einzelner Proben mit individuell anpassbaren Genregionen, die einfach über eine Textdatei festgelegt werden können. Die Pipeline habe ich mit 252 Proben – einschließlich archivierten Kryogewebes und diagnostischen Schnellschnitten – entwickelt und validiert. Hierbei kamen Ad-hoc-Modelle für die Methylierungsklassifikation und den Nachweis der MGMT-Promoter-Methylierung zum Einsatz. Für die Vorverarbeitung, Variantendetektion und -annotation nutzte ich hochmoderne, öffentlich zugängliche Tools und entwickelte Strategien zur Beschleunigung der Datenverarbeitung. Zusätzlich zeige ich das Potenzial der Pipeline auf, intraoperative Ergebnisse zu liefern: Anhand von 18 Proben aus zwei unabhängigen Zentren wird demonstriert, dass Rapid-CNS^2 innerhalb eines 30-minütigen intraoperativen Zeitfensters Echtzeit-Methylierungsklassifikationen und DNA-Kopienzahlberichte bereitstellt. Das vollständige molekulare Profiling erfolgt dann innerhalb von 24 Stunden, was das gesamte Spektrum molekularer Veränderungen abdeckt, die für die Diagnose und gezielte Therapie von ZNS-Tumorsubtypen von Bedeutung sind. Damit verkürzt Rapid-CNS^2 die wochenlange Bearbeitungszeit herkömmlicher Methoden erheblich.

Um den Zugang zum MNP-Klassifikator weiter zu erleichtern, entwickelte ich MNP-Flex, eine plattformunabhängige Version des MNP-Klassifikators, die 184 ZNS-Tumorklassen abdeckt.

MNP-Flex wurde an einer globalen Kohorte von über 78.000 Proben validiert, darunter gefrorene sowie formalinfixierte, paraffineingebettete (FFPE) Proben, die mit fünf verschiedenen Methylierungsprofilierungstechnologien verarbeitet wurden. Mit klinisch relevanten Schwellenwerten erzielte MNP-Flex eine beeindruckende Genauigkeit von 99,6 % für Methylierungsfamilien und 99,2 % für Methylierungsklassen.

Gemeinsam bieten Rapid-CNS^2 und MNP-Flex einen umfassenden Workflow für die ZNS-Tumordiagnostik. Rapid-CNS^2 ermöglicht intraoperative Echtzeitberichte über Methylierungsklassifikationen und Kopienzahlveränderungen, die die chirurgische Strategie unterstützen, während das vollständige molekulare Profil und die detaillierte Methylierungsklassifikation durch MNP-Flex bereits am Folgetag bereitstehen, um klinische Entscheidungen und therapeutische Maßnahmen zu erleichtern. Dieser Workflow ist nicht nur kosteneffektiv und kompakt in der Anwendung, sondern verwendet auch leicht zugängliche Labor- und Bioinformatik-Tools. Rapid-CNS^2 ist auf GitHub verfügbar, und MNP-Flex kann als Webservice für Forschungszwecke unter https://mnp-flex.org abgerufen werden. Dieser integrierte Ansatz optimiert die molekulare Diagnostik von ZNS-Tumoren, erweitert den globalen Zugang zu präzisen, molekular fundierten Klassifikationen und trägt letztlich zur Verbesserung der Patientenergebnisse bei.

Acknowledgements

As the saying goes, "The journey is the reward." This PhD has been as much about the path taken as the final destination, and I owe a profound debt of gratitude to those who have walked this road with me, offering their support, guidance, and encouragement along the way.

First and foremost, I would like to express my gratitude to my supervisor, mentor, guide and examiner Prof Dr Dr med. Felix Sahm. Thank you for giving me the opportunity to pursue my PhD working on some of the most cutting-edge translational topics. Your support has extended far beyond academic guidance and I am grateful for the countless opportunities you have provided me with throughout the course of this work- from collaborating with leading experts and presenting my research at multiple conferences to being exposed to various career-building experiences, your mentorship has been invaluable. Your ideas, encouragement, and belief in my potential have shaped not only this thesis but also the trajectory of my career.

I would like to thank my co-supervisors Prof Matthias Schlesner and Prof Moritz Gerstung for so effortlessly embracing me into their labs. To Matthias, I'm especially grateful for your calm and composed guidance throughout. Your advice on bioinformatics and beyond, has been invaluable. I truly appreciated the BODA BCT breaks during the pandemic—they provided much-needed moments of connection and support during those challenging times. Your calm approach made everything seem manageable, and I'm thankful for your wisdom and for being a steady presence throughout my PhD. To Moritz, thank you for adopting me (literally) into your lab and for being a fantastic source of advice, scientific curiosity and bizarre facts. Your ability to ask the right questions has guided me through complex problems. While this thesis marks the completion of one chapter, I'm excited that our work together is only just beginning, with many more projects and product ideas ahead.

I would like to express my thanks to Prof Benedikt Brors for agreeing to be my examiner and member of my thesis advisory committee. Your guidance has been truly valuable. Many thanks to Dr Christiane Opitz and Dr Pei-Wei Chi for agreeing to be examiners for my thesis.

I would like to extend my sincere thanks to my thesis advisory committee- Prof Benedikt Brors, Prof Felix Sahm, Prof Matthias Schlesner, Dr Violaine Goidts, Prof David Jones and Dr Sophie Weil for their invaluable feedback and guidance which have significantly shaped this work.

I would like to thank all members of the Department of Neuropathology, Sahm lab, BODA and AI in Oncology groups for being excellent colleagues and friends.

I want to thank Prof. Andreas von Deimling for his leadership, infectious excitement for research and constant encouragement. Your commitment to research has always been a source of inspiration. I would also like to express my gratitude to Prof. Stefan Pfister for his support and faith in my abilities.

I want to thank all my collaborators particularly Prof Matthew Loose, his lab, Dr Simon Paine, Prof Einar Vik-Mo and his group for their collaboration and excellent conversations that always made me more excited about science.

I would like to express my gratitude to Kirsten Göbel for being a friend and my favourite person to work on nanopore sequencing with. I could not have managed a lot of this work without you. Many thanks to Dr Helin Dogan for her friendship and dedication to initially setting up this work.

Many thanks to Dr Martin Sill and Dr Daniel Schrimpf for their technical and moral support, your guidance has been a life-saver.

To the friends who became family- Zaira, Dina, Micha, Christina and Enrique, thank you for being there for me always, through all my conundrums, I will always cherish our time together.

Finally, and most importantly, much gratitude to my family for always being there for me and supporting me in all my unreasonable endeavours. Mum, Dad, Noorain, Rushda and Zoha- thank you for always believing in me.

Contents

List of Figures

List of Tables

Nomenclature

5mC: 5-Methylcytosine

6mA: N6-Methyladenine

AUC: Area under the curve

BAM: Binary Alignment/Map format for storing aligned sequencing data

bp: Base Pairs

CIMP: CpG Island Methylator Phenotype

cIMPACT-NOW: Consortium to Inform Molecular and Practical Approaches to CNS Tumor Taxonomy- Not Official WHO

cDNA: Complementary DNA

CNS: Central Nervous System

CNV: Copy number variation

CPU: Central Processing Unit

ddNTPs: Dideoxynucleotides

DKFZ: German Cancer Research Center

DNA: Deoxyribonucleic Acid

EPIC: Infinium HumanMethylationEPIC BeadChip

FAST5/ POD5: File formats for storing raw nanopore sequencing data

FISH: Fluorescence In-Situ Hybridisation

GB: Gigabytes

Gb: Gigabase

GPU: Graphical Processing Unit

HMM: Hidden Markov Model

Indel: Insertion / Deletion

KB: Kilobytes

Kb: Kilobase

KiTZ: Hopp Children's Cancer Center Heidelberg

MB: Megabytes

Mb: Megabase

MinION/ GridION/ PromethION: Devices from Oxford Nanopore Technologies

MMR: Mismatch Repair

MNP: Molecular Neuropathology methylation classifier

NGS: Next Generation Sequencing

ONT: Oxford Nanopore Technologies

PCR: Polymerase Chain Reaction

RNA: Ribonucleic acid

RF: Random Forest

SBS: Sequencing by Synthesis

SBL: Sequencing by Ligation

SMRT: Single Molecule Real-Time

SNV: Single Nucleotide Variant

SV: Structural Variant

TB: Terabytes

TCGA: The Cancer Genome Atlas

TMB: Tumour Mutational Burden

t-SNE: t-distributed Stochastic Neighbour Embedding

UMAP: Uniform Manifold Approximation and Projection

UKHD: University Hospital Heidelberg

WES: Whole Exome Sequencing

WHO: World Health Organisation

WGBS: Whole Genome Bisulphite Sequencing

WGS: Whole Genome Sequencing

XGBoost: Gradient boosting library

About this thesis

This thesis marks the commencement of my overarching quest to enable access to modern precision cancer diagnostics and subsequently therapy to one and all irrespective of their geographical location, socioeconomic status, ethnicity or gender. Coming from a developing country, I have first-hand observed the disparity in care and treatment among my countrymen primarily owing to inequality of access. It is this inequality that has fuelled my determination to bridge the gap between advanced scientific innovations and the people who need them the most. Through this work, I aim to contribute to the democratisation of healthcare by developing methods that are not only scientifically robust but also accessible and affordable for implementation in diverse settings, particularly in under-resourced regions. I firmly believe that precision diagnostics should not be a privilege limited to the affluent or those in developed nations, but a basic healthcare right available to every individual, no matter where they are or who they are. This thesis represents my first step towards realising that vision.

In this thesis, I commence with a comprehensive literature review of CNS tumours, their diagnosis, and an overview of how evolution of sequencing technologies has impacted advances in diagnostics and treatment approaches. This sets the stage for the development and application of new methodologies in the field.

The core of my work is presented in two chapters, each dedicated to a specific method I have developed- Rapid-CNS^2 and MNP-Flex. In these chapters, I first provide a primer on the technology underlying these methods, along with a discussion of the work of giants that I have had the privilege of standing on the shoulders of. Next comes a detailed explanation of the methods, including how I validated them detailing experiments and analyses. I follow this with a results section that states and discusses individual results. The chapters conclude with a broad discussion of implications of the methods, developing research in the field, along with reflections on how these methods could potentially be improved in future work.

Finally, I conclude with a general discussion and outlook chapter. This chapter provides a broader context for my work, exploring its potential impact on the field while acknowledging the limitations of the study. It also looks ahead to possible future developments and refinements that could build on the foundation that I have attempted to establish.

Disclosures

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text. I have used ChatGPT to proofread and edit descriptive sections of this thesis. I used DALL-E to create the logos for the tools developed in this thesis. I created the analyses figures in R and edited them in Affinity Designer. I created the descriptive figures using the professional license of Biorender provided by the Hopp's Children's Cancer Center (KiTZ). I have referenced the figures taken from previously published work or web portals. Where applicable, I have taken permission from the publisher to reprint the work.

Areeba Jamilkhan Patel

2024

Chapter 1 Introduction

1.1 Sequencing technologies

The ability to sequence DNA has revolutionised biology and medicine, providing deep insights into the genetic blueprint of organisms. The development and accessibility of advanced sequencing technologies have been pivotal in ushering in a molecular era in CNS tumour diagnostics and treatment. As these technologies evolved, they allowed for the generation of large-scale molecular data that was previously unattainable, making it possible to identify distinct tumour subtypes and actionable alterations in larger populations. This influx of data provided a foundation for refining diagnostic criteria and testing new hypotheses, leading to more accurate tumour stratification and the development of targeted therapeutic approaches that were not feasible before. Sequencing technology has undergone significant evolution since the first DNA sequencing methods were developed, transitioning from laborious, manual techniques to highly automated, high-throughput platforms (**Figure 1-1**). Early methods, such as Sanger sequencing, paved the way for massive genomic projects like the Human Genome Project, while next-generation sequencing (NGS) introduced massively parallel sequencing, drastically reduced costs and time. The latest innovations in third-generation sequencing now offer longer reads and single-molecule accuracy, opening new avenues for studying complex genomes, structural variants, and epigenetic modifications leading to the complete telomere-to-telomere assembly of the human genome. In this section, I give an overview of the progression of sequencing technologies from their inception to the current state of the art while simultaneously commenting on their impact on advances in modern CNS tumour diagnostics.

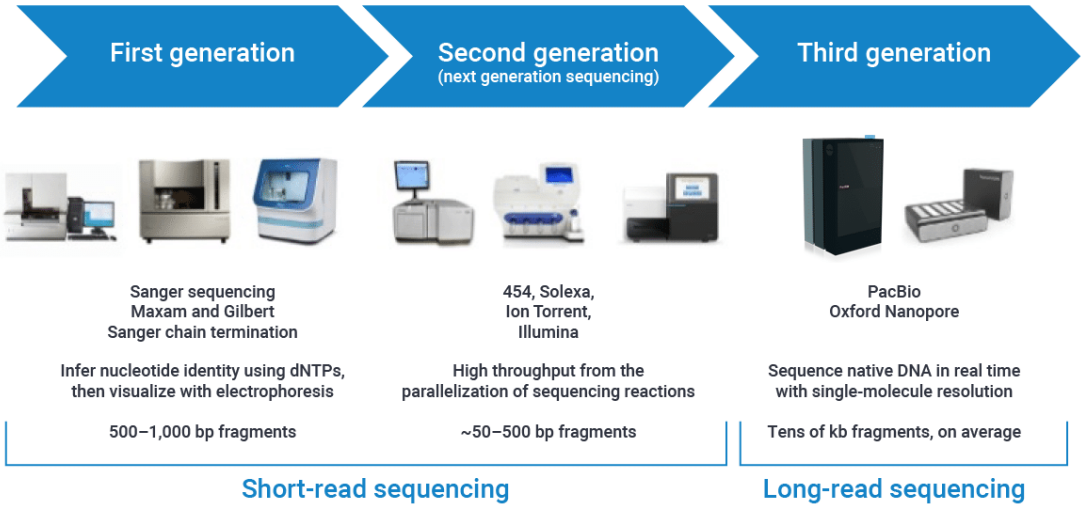


Figure 1-1 Evolution of sequencing technologies (Reprinted from https://www.pacb.com/blog/the-evolution-of-dna-sequencing-tools/, accessed on 16th Sep 2024)

1.1.1 First generation sequencing: the genesis of genomics

The journey towards understanding the genetic code began with the discovery of the structure of DNA in 1953, when Watson and Crick unveiled their iconic double helix shape at the Eagle pub in Cambridge, building on crystallographic data from Rosalind Franklin and Maurice Wilkins[1,2]. This ground-breaking discovery laid the foundation for understanding DNA replication and its role in encoding proteins. However, there was no technology to actually "read" or sequence DNA. The landscape of DNA sequencing changed drastically with two influential developments in the mid-1970s: the "plus and minus" system developed by Alan Coulson and Fred Sanger, and the chemical cleavage method of Allan Maxam and Walter Gilbert[3]. This technique led to the sequencing of the first DNA genome, bacteriophage φX174, which remains a reference genome for many modern sequencing labs. Despite its added complexity, Maxam-Gilbert sequencing became the first widely adopted method for DNA sequencing.

The most significant breakthrough in first-generation sequencing came in 1977, when Sanger developed the chain-termination method, also known as Sanger sequencing[4]. This technique relied on the incorporation of dideoxynucleotides (ddNTPs), which lacked the 3' hydroxyl group required for DNA chain elongation. By mixing ddNTPs with regular nucleotides in a DNA polymerase reaction, the sequencing process was halted at random points, generating

fragments of varying lengths. These fragments were then separated on polyacrylamide gels, and the resulting bands were used to infer the DNA sequence via autoradiography.

Sanger sequencing had a profound impact on the field of genomics and was particularly transformative for fields such as oncology, where understanding the genetic basis of diseases like cancer became crucial. By enabling the sequencing of oncogenes, Sanger sequencing laid the groundwork for identifying key genetic mutations involved in tumour formation and progression. For instance, the identification of mutations in genes like *IDH1* and 1p/19q co-deletions in gliomas, a crucial component in the diagnosis and treatment planning of CNS tumours, was made possible through the early use of Sanger sequencing.

Improvements in Sanger sequencing followed in the late 1980s and 1990s, including the shift to fluorescent labelling and introduction of capillary electrophoresis, enabling the automation of DNA sequencing[5]. Sanger sequencing was also instrumental in the Human Genome Project, which produced the first draft of the human genome ahead of schedule[6,7]. The introduction of polymerase chain reaction (PCR) and recombinant DNA technologies further accelerated genomic research by providing ample quantities of high-purity DNA for sequencing.

These early technologies, were instrumental in shaping the future of CNS tumour diagnostics and treatment, providing insights that continue to inform current therapeutic strategies. Even today, automated Sanger sequencing is the gold standard for mutation detection for molecular diagnostics and is widely used by modern labs to test specific alterations.

1.1.2 Second (next) generation sequencing

The advent of next-generation sequencing (NGS) marked a paradigm shift in genetics by enabling DNA sequencing at unprecedented speed. Unlike first-generation sequencing methods, which relied on labour-intensive processes like Sanger sequencing, NGS leveraged massively parallel sequencing to generate vast amounts of data in a fraction of the time and cost. The first major breakthrough in NGS came with the introduction of pyrosequencing, developed by Pål Nyrén and colleagues[8,9]. The method measured the release of pyrophosphate during nucleotide incorporation, which was subsequently converted into light through an enzymatic reaction involving ATP sulfurylase and luciferase. However, it had limitations, particularly in handling homopolymer regions, where the signal became noisy and

difficult to interpret beyond a stretch of four or five identical nucleotides. 454 Life Sciences, founded by Jonathan Rothberg, was the first to commercialise pyrosequencing, producing the 454 GS 20 system. This enabled high-throughput sequencing for the first time. The platform made it possible to sequence entire genomes, famously of DNA structure co-discoverer James Watson, at a significantly lower cost than traditional Sanger sequencing. Nevertheless, challenges with interpreting homopolymer regions and the cost of reagents led to other NGS technologies, particularly Illumina's platform, to dominate the market.

Illumina utilised the sequencing-by-synthesis (SBS) platform which included incorporation of fluorescently labelled nucleotides to synthesize complementary DNA strands, one nucleotide at a time. These nucleotides are modified with reversible terminator groups, which blocks further synthesis after the addition of each nucleotide, ensuring high accuracy by allowing only one base to be added per cycle. After each nucleotide incorporation, the terminator and fluorophore are cleaved, to begin the next cycle of synthesis. SBS operates through the generation of short reads, typically ranging between 50–300 base pairs (bp) in length. A key component of Illumina sequencing is the generation of clonal DNA clusters through a process called solid-phase bridge amplification. DNA is fragmented, followed by adapter ligation. These adapters enable binding to bind to complementary primers immobilised on a solid surface, such as a flow cell. DNA polymerase then synthesises a complementary strand, forming a bridge-like structure. Through repeated cycles, millions of identical copies of each DNA fragment are created, forming dense clusters of DNA molecules on the flow cell. In recent Illumina platforms, such as the NovaSeq, the use of patterned flow cells has optimised this process. The NovaSeq 6000 is capable of generating over 3 terabytes (TB) of data in a single run, enabling population-scale studies of human genomes at 30X coverage in a matter of days.

While Illumina's SBS dominates short-read sequencing, sequencing by ligation (SBL) is another prominent method used by some platforms, such as Thermo Fisher's SOLiD platform. SBL operates by hybridising short fluorescently labelled oligonucleotide probes to the DNA template. These probes contain known bases at specific positions, which are ligated to adjacent oligonucleotides if they are complementary to the template. The emission spectra from the fluorophores are detected to identify the incorporated nucleotides. A major advantage of SBL is its high basecalling accuracy (~ 99.99). However, the short-read lengths produced by SBL, typically between 50 and 75 bp, limit its utility in applications such as genome assembly or detection of structural variants.

Moore's law, which predicts the doubling of computing power every two years, has traditionally guided expectations for technological advancement in fields reliant on computation and data processing. The cost of sequencing has decreased at a rate surpassing Moore's law, with the price of sequencing a human genome plummeting from nearly $3 billion during the Human Genome Project to approximately $1,000 today (**Figure 1-2**). The shift to massively parallel sequencing, automation, miniaturisation of reaction volumes, and increased throughput from platforms like the NovaSeq have drastically reduced sequencing costs. Additionally, bioinformatics innovations have streamlined data processing, further lowering expenses.

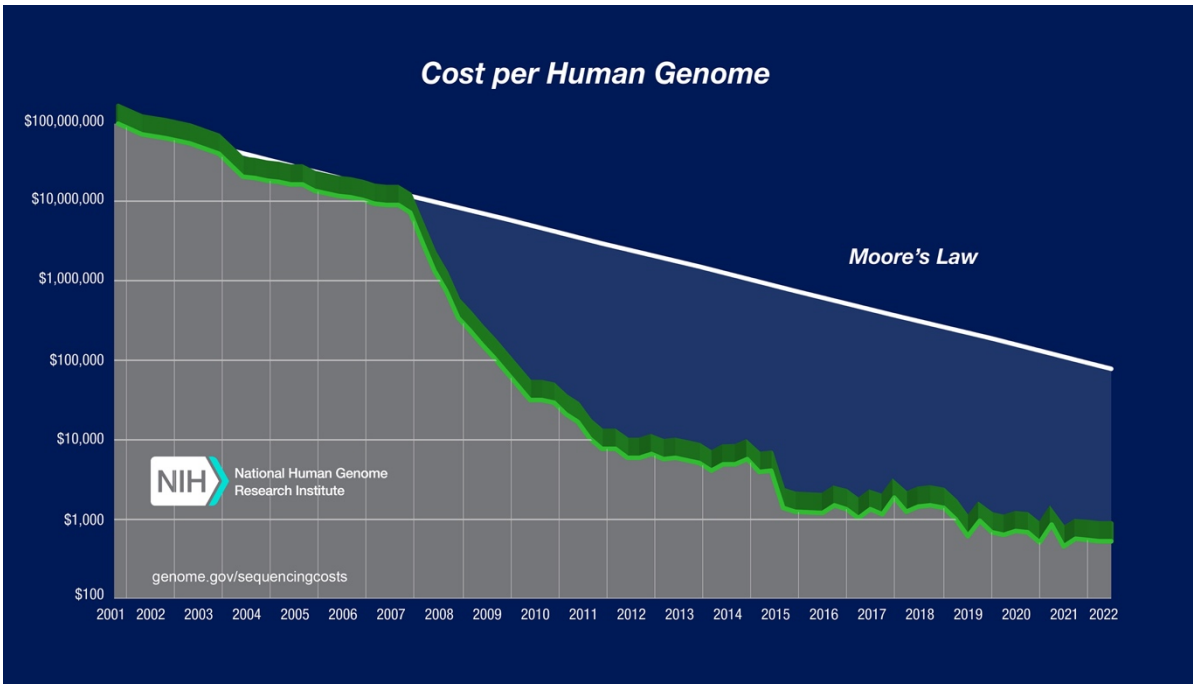


Figure 1-2 Cost per human genome over time (Reprinted from NHGRI website https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data, accessed on 28th September 2024)

Despite the dramatic reduction in cost of sequencing, whole-genome sequencing (WGS) remains a costly endeavour, particularly for large-scale studies. A recent study has shown that although WGS results in an estimated fivefold increase in the total number of assayed variants over WES + array genotyping with imputation, the number of detected signals differed by only 1% for both SNVs and gene-based association analyses[10]. To address this, targeted sequencing strategies were developed to focus on specific regions of interest within the genome. Targeted sequencing involves identification of relevant genomic regions, design and ordering of primers, fragmentation, target enrichment by either hybrid-capture or PCR amplicons, sequencing and finally data analysis of targeted regions. A combination of WGS,

WES and NGS panel sequencing of CNS tumours has resulted in the accumulation of large population-level datasets across the globe. The boom in accessibility of NGS caused rapid identification of new molecular markers by consortiums such as The Cancer Genome Atlas (TCGA) in 2015, leading to addition of molecular markers for the first time to the WHO classification of CNS tumours (**Figure 1-3**).

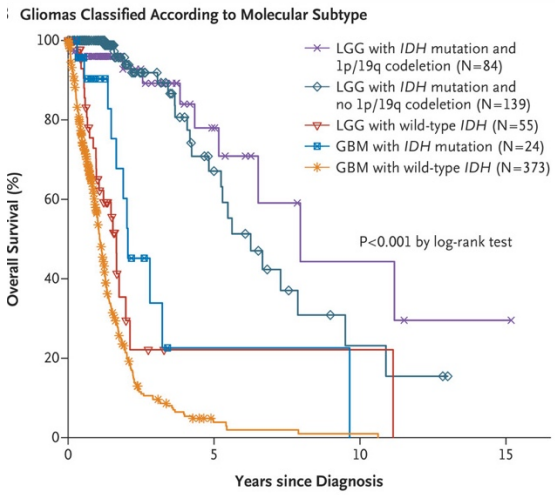


Figure 1-3 Clinical outcomes of diffuse glioma stratified by molecular subtypes (Reproduced with permission from The Cancer Genome Atlas Research Network 2015[11], Copyright Massachusetts Medical Society.)

NGS also enabled epigenetic profiling of whole genomes using bisulphite conversion called whole genome bisulphite sequencing (WGBS)[12]. WGBS and similar epigenetic assays resulted in finally understanding the mutationally-cold tumours of the CNS especially those with no obvious driver alteration[13-15].

In parallel, Illumina developed a series of DNA methylation arrays to offer more accessible alternatives to WGBS, with increasing coverage of CpG sites across the human genome. The initial HumanMethylation27 BeadChip (27K) covered over 27,000 CpGs, followed by the HumanMethylation450 BeadChip (450K), which expanded coverage to more than 450,000 CpGs[16,17]. In 2016, the HumanMethylationEPIC BeadChip (850K) was introduced, covering over 850,000 CpGs and extending coverage to include regulatory elements like enhancers and CTCF-binding sites, which play critical roles in transcriptional regulation[18]. Each chip can profile 8 samples at a time. These scalable solutions were key in the development of the MNP methylation classifier that has positively transformed CNS tumour diagnostics[19]. The most recent version, released in 2023, the Infinium HumanMethylationEPIC v2.0 BeadChip (900K), further expands this to cover 900,000 CpG sites, incorporating regions identified by major projects like ENCODE and FANTOM5[20-22]. This includes 200,000 new CpGs in open chromatin and enhancer regions, critical for understanding the influence of distal regulatory elements on gene expression.

Despite these innovations and scalability, the reliance of NGS on short reads (50–300 bp) makes it difficult to accurately sequence complex genomic regions like repetitive sequences and structural variants. Additionally, NGS requires DNA amplification, introducing biases that prevent the detection of base modifications, such as DNA methylation, which necessitates separate, indirect methods like bisulphite treatment. Third-generation sequencing technologies, such as those by PacBio and Oxford Nanopore, address these issues by producing much longer reads to the order of a few Mb and sequencing single molecules without amplification.

1.1.3 Third generation sequencing

Over the past thirty years, the field of single-molecule detection has caught speed giving rise to third-generation sequencing. Among these innovations, nanopore-based sequencing has gained particular attention due to its compact devices and simplified library preparation, allowing researchers to sequence individual DNA or RNA molecules without amplification. The early concepts of nanopore sequencing date back to the 1980s, when independent laboratories, including those led by David Deamer, George Church, and Hagan Bayley, postulated the feasibility of using nanopores for single-molecule detection[23-25]. This idea hinged on the principle that nucleic acids could be driven through a nanoscale pore by an electric current, with each nucleotide base causing a characteristic disruption in the current as it passed through. In 2003, Hagan Bayley became a professor of chemical biology at the University of Oxford and co-founded Oxford Nanopore Technologies (ONT) in 2005 with a team experienced in biotech and diagnostics. ONT licensed DNA sequencing patents in 2008 and focused on strand sequencing, leading to the MinION device, unveiled by CTO Clive Brown at the 2012 AGBT Meeting. In 2014, ONT launched the MinION Access Program, distributing devices for large-scale collaboration. Nanopore sequencing technology has matured significantly since these early breakthroughs. The initial experiments used the staphylococcal alpha-haemolysin (αHL) nanopore for nucleic acid translocation. ONT has released eight versions of the nanopore and motor protein so far starting from R6 in 2014. A significant advancement came with the R9 version, which utilised the Curlin sigma S-dependent growth subunit G (CsgG) from *Escherichia coli*, resulting in improved sequencing accuracy (~87%, compared to ~64% in R7) and faster translocation speeds (~250 bases per second versus ~70 bases per second in R7). The R10 and R10.3 nanopores feature two sensing regions (or reader heads) to improve homopolymer sequencing accuracy.

This technology employed by ONT has made this technology widely accessible by offering long-read sequencing capabilities at competitive costs. Unlike other platforms, nanopore sequencing enables sequencing of entire genomic regions in a single read, making it especially valuable for detecting structural variants, exploring epigenetic modifications, and conducting molecular biomarker discovery. In nanopore sequencing, a biological or synthetic nanopore is embedded in a membrane separating two chambers filled with electrolytic fluid, typically KCl or Ag/AgCl systems. The sequencing chamber holds the nucleic acid molecules, while the trans side receives them after they pass through the pore. A voltage bias applied across the membrane causes an ionic current to flow through the nanopore, which is monitored by a patch-clamp amplifier or, in the case of those developed by Oxford Nanopore Technologies (ONT), by compact ASIC chips. Since this current disruption is directly affected by the structure of the base, it follows that base modifications can also be detected by sensitive basecalling techniques. The negatively charged DNA or RNA molecules are drawn towards the anode and thread through the nanopore. As they translocate through the pore, each base disrupts the ionic current in a uniquely, producing a pattern that can be used to infer the sequence. Two key parameters are used to characterise the translocation process: the event duration, which refers to the time a molecule spends in the pore, and the magnitude of the current blockade, which corresponds to the interaction between the nucleotides and the ionic current. These disruptions, or nucleotide fingerprints, are then mapped back to reveal the sequence of the molecule being analysed.

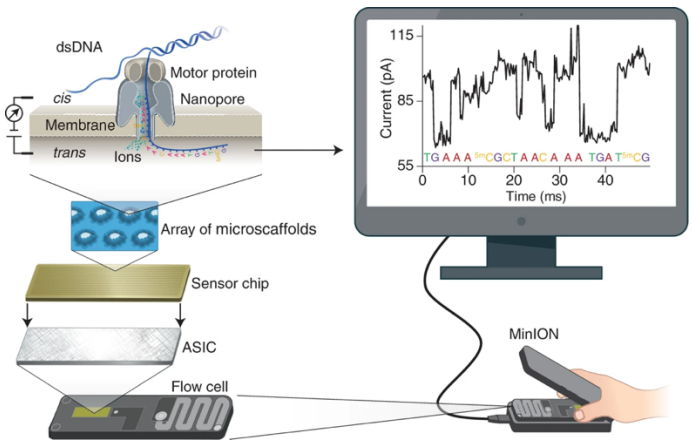


Figure 1-4 Mechanism of nanopore sequencing on a portable MinION (Reprinted with permission from Wang et. al 2021[26])

Figure 1-4 demonstrates the mechanism of sequencing and basecalling using the handheld device by ONT- the MinION[27]. As evident from the figure, subtle changes in the current are observed due to base modifications that are represented in a 'squiggle'. This squiggle is translated to base space using basecallers that have evolved from employing statistical tests, hidden Markov models (HMM), neural networks and recently transformer-based models[28-35].

Among third-generation sequencers, ONT and PacBio are the two dominant players. PacBio has the single-molecule real-time (SMRT) sequencing technology that uses fluorescently labelled nucleotides to track DNA replication in real-time. It is particularly known for its high accuracy, especially with the advent of HiFi (high-fidelity) sequencing, which repeatedly sequences the same strand of DNA to produce a highly accurate consensus read. However, the need for multiple passes through the same sequence limits the length of reads, typically to around 20 kbp. ONT's approach, on the other hand, detects changes in ionic current as nucleic acids pass through a nanopore, rather than relying on fluorescence. This technology has evolved rapidly, reducing the error rates historically associated with nanopore sequencing, which once ranged as high as 38%. Current error rates for ONT are now on par with PacBio, owing to improvements in basecalling algorithms and library preparation methods. ONT's ability to generate ultra-long reads—sometimes extending beyond a megabase—offers a distinct advantage in sequencing highly repetitive or complex regions that are difficult to resolve using PacBio or short-read platforms. In line with expectations, nanopore sequencing has been instrumental in closing gaps in the genomes by accurately tracing hard to map regions of the genome finally leading to completion of whole chromosome assemblies[36,37].

ONT devices are user-friendly and accessible, as they do not require advanced computing resources or expertise for basic data analysis, making them practical for many laboratories. There are four major device families- Flongle, MinION, GridION and PromethION and two flowcell types- MinION and PromethION. MinION flowcells are smaller with 2,048 pores and 512 channels, while PromethION flowcells have 12,000 pores and 2,675 channels. MinION is the smallest handheld device that can use one flowcell at a time. the GridION is a multiplexed version with 5 MinION flowcells and integrated compute. Currently, there are 4 PromethION devices- PromethION24, PromethION48, PromethION2 Solo (P2 Solo) and PromethION2i (P2i) which is with integrated compute. An overview of all devices is provided in **Figure 1-5**.

Flow cell name	Flongle	MinION Flow Cell	PromethION Flow Cell
Number of channels	126	512	2,675
Theoretical maximum output*	2.8 Gbases	50 Gbases	290 Gbases

Device name	MinION	MinION Mk1C	GridION	PromethION 2 Solo	PromethION 2	PromethION 24	PromethION 48
Flow cell compatibility	Flongle, MinION			PromethION			
Number of flow cells that can be run	1	1	5	2	2	24	48

Figure 1-5 Overview of ONT devices (Reprinted with permission from Pugh et. al 2023[38])

The sequencing is controlled using MinKNOW, the operating software for ONT devices, which manages sequencing parameters, tracks samples, and handles real-time analysis. With integrated basecallers, MinKNOW can also performs base calling, thus converting the raw FAST5 or recently the POD5 files to FASTQ and BAM files. FAST5 and POD5 files are organised in a nested format for easy extraction of specific information. With current basecalling tools like guppy and dorado, BAM files also indicate modified base information depending on the model employed.

Nanopore sequencing enables the direct detection of DNA and RNA modifications through the analysis of current shifts. Tools such as Nanopolish, and DeepSignal have been developed to detect modifications like 5mC and 6mA in DNA. ONT is particularly effective in detecting 5mC with high accuracy at the single-molecule level. RNA modification detection has also advanced, with tools identifying modifications such as m5C, m6A and pseU in RNA. However, the detection of more RNA modifications with single-nucleotide resolution is still under development.

Initial studies have demonstrated effective fusion detection, short tandem repeat detection and allele-specific chromosome arm telomere length measurement[39-42]. Nanopore sequencing is

currently the only platform that allows direct RNA sequencing opening up avenues for detecting RNA modifications without manipulations like chemical conversions[43,44]. Long reads have also enabled early studies into single cell RNA sequencing where specific transcript isoforms as well as fusions can be resolved on a single cell and spatial level[45,46]. Recent studies have also demonstrated the feasibility of sequencing single protein molecules in their native form using nanopores[47-49]. With further improvements in the technology, it follows that these will undergo considerable developments leading to better detection strategies and deeper understanding of cancer and beyond.

1.2 Cancer

Cancer is the second leading cause of death in the world accounting for nearly 10 million deaths in 2020. The number of new cancer cases is expected to increase to 29.9 million per year by 2040, and the number of cancer-related deaths is expected to increase to 15.3 million[50]. Nevertheless, cancer mortality has been steadily decreasing across the globe. In the United States, for example, cancer death rates fell by 33% from 1991 to 2020, preventing an estimated 3.8 million deaths. This drop is largely attributed to nationwide screening programs, public health initiatives like tobacco control and development of targeted therapeutic approaches driven by precise identification of molecular markers through cutting-edge diagnostic methods.

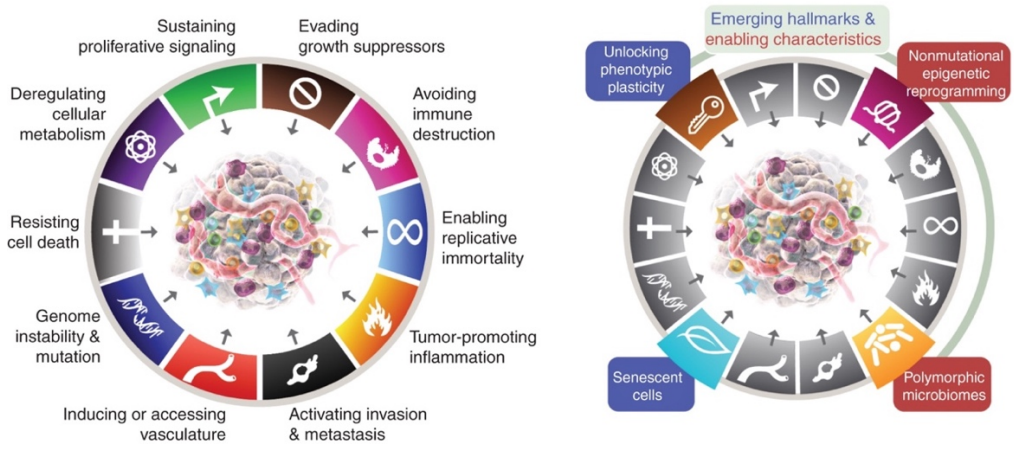


Figure 1-6 New hallmarks of cancer (Reprinted with permission from Hanahan et. al 2022[51])

The key capabilities of cancer cells during oncogenesis and progression have been famously described by the hallmarks of cancer elucidated by Douglas Hanahan and Robert Weinberg[52].

These hallmarks include sustaining proliferative signalling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion and metastasis. These fundamental characteristics allow cancer cells to grow uncontrollably and spread throughout the body, overcoming the normal regulatory mechanisms of the cellular environment. In the recent update as shown in **Figure 1-6**, emerging hallmarks have been proposed, such as non-mutational epigenetic reprogramming, further expanding our understanding of the adaptability of cancerous cells[53].

Molecular diagnostics plays a vital role in understanding and identifying the hallmarks of cancer, offering precise tools to detect genetic and epigenetic alterations that drive tumour growth and progression. The hallmarks of cancer, such as sustaining proliferative signalling or evading growth suppressors, are rooted in molecular changes that are detectable through modern diagnostic technologies. For example, DNA sequencing can reveal mutations in oncogenes or tumour suppressor genes, which are key to understanding how cancer cells bypass normal regulatory pathways. Epigenetic changes, particularly DNA methylation patterns, are another crucial layer in cancer biology reflecting the emerging hallmark. These alterations can silence or activate genes involved in hallmarks like resisting cell death or enabling replicative immortality. Methylation profiling, as part of molecular diagnostics, is now used to classify tumour types with greater precision, often when histopathology alone cannot provide clear answers. Tools like nanopore sequencing and methylation arrays allow for the detection of aberrant epigenetic modifications that are associated with specific cancer types, leading to more tailored treatments.

Through its multi-omic character, molecular diagnostics can be instrumental in tracking how cancers evolve, revealing clonal changes that contribute to metastasis, therapeutic resistance, or immune evasion. This adaptability of tumours, reflected in changes to the molecular landscape, is a key aspect of the hallmarks, and molecular diagnostics enables continuous monitoring, particularly in recurrent or aggressive cancers. Overall, by targeting these molecular changes associated with the hallmarks of cancer, diagnostics guide personalised medicine approaches, improve patient outcomes, and allow for more targeted therapies that disrupt the fundamental biological processes of cancer.

There has been a substantial shift from morphology to molecular analyses in tumour diagnostics driven by advances in technologies like NGS, methylome profiling, and proteomics, offering granular classification and better treatment insights. Methylation analysis

is key for determining tumour lineage, while NGS focuses on tumour-specific alterations like gene fusions and mutations, especially in genetically simple paediatric tumours. Proteomics, particularly through mass spectrometry, holds promise for understanding cellular signalling pathways to guide targeted therapies. The integrated, layered diagnostic system of the WHO seeks to standardise diagnoses globally, incorporating molecular data, tumour grading, and morphologic features for a comprehensive evaluation.

1.2.1 Central nervous system tumours

CNS tumours represent around 1.5% of all cancers worldwide, with approximately 308,000 new cases diagnosed annually. CNS tumours are the second most common cancer in children (after leukaemia) and the most prevalent solid tumour in children accounting for around 20-25% of paediatric cancers[54]. Annually, around 241,000 people worldwide die from brain and CNS tumours[55]. In 2017, the 5-year survival rate for primary malignant brain tumours was a bleak 36%, nevertheless marking a 5% improvement since 1997[50]. This increase is largely due to advances in molecular diagnostics, which have identified molecular targets and enabled more effective therapeutic strategies.

For over a century, CNS tumours have been diagnosed based on histopathology. This has been reflected in the WHO guidelines for CNS tumour classification. In 2021, the fifth edition of the WHO Classification of Tumours of the Central Nervous System (WHO CNS5) represented a paradigm shift in the classification of CNS tumours, with molecular testing taking the centre stage in diagnosis and classification[56-58]. Issuing an integrated diagnosis thus involves multiple layers of information, beginning with conventional histopathological features enhanced by molecular markers. An important example of this shift is the refined grading and nomenclature for diffuse gliomas. The diagnostic flowchart for such adult and paediatric diffuse gliomas is demonstrated in **Figure 1-7**. The schematic demonstrates the need for multiple molecular testing to make a WHO compatible integrated diagnosis.

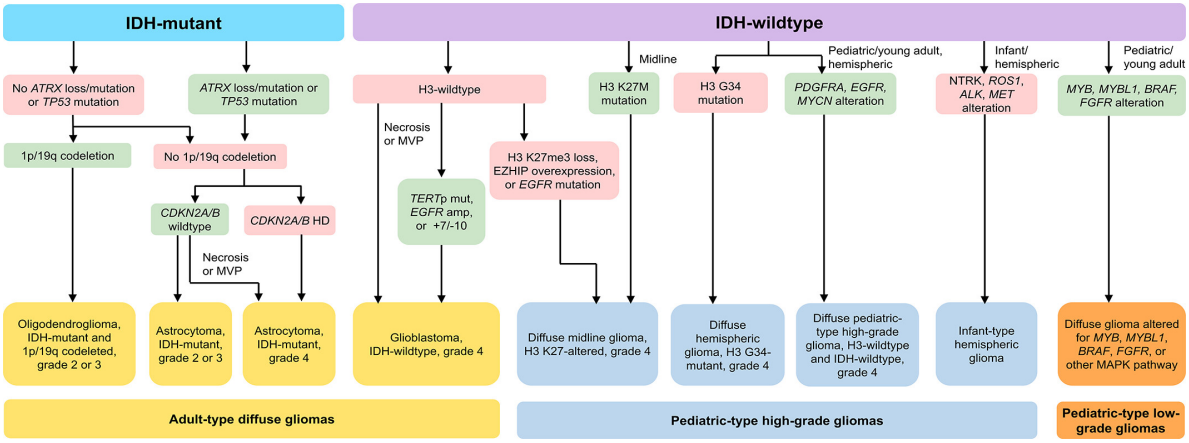


Figure 1-7 Diagnostic flowchart for diffuse gliomas. (Reprinted with permission from Park et. al 2023[59])

The traditional reliance on histological features is now augmented by molecular markers like mutations, copy number alterations, gene fusions and methylation-based classification, which are increasingly recognised as defining characteristics for many tumour types (**Figure 1-8**). Molecular profiling can either assist or independently classify a tumour type, improve prognostic accuracy by supplementing histology-based tumour grading, and identify potential therapeutic targets for personalised treatment. These layers are often interconnected, requiring a strong understanding of how to interpret each molecular marker within the relevant diagnostic context for proper clinical application. Childhood tumours, though heterogeneous and less common than adult cancers, have distinct origins and lower genetic complexity, often driven by a single clonal event such as an oncogenic fusion. Unlike adult tumours, which are frequently linked to long-term exposure to carcinogens, paediatric tumours typically arise from immature cell types with a block in normal development. For the first time, paediatric tumours have been assigned a separate volume in the new WHO classification system[60].

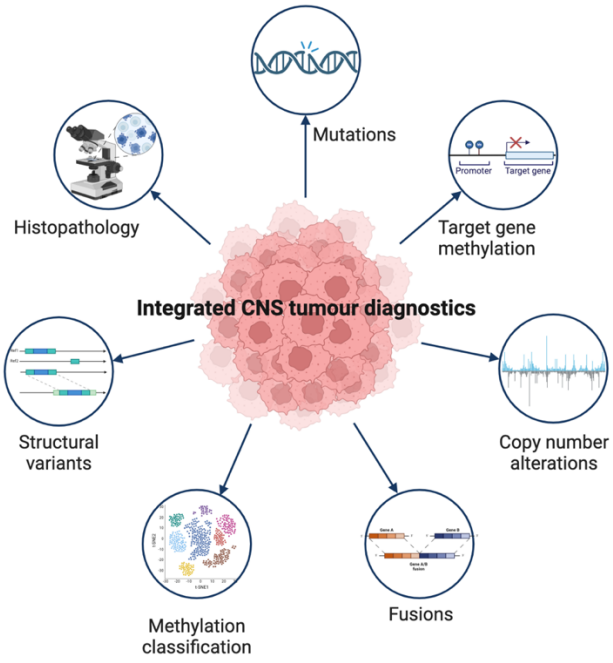


Figure 1-8 WHO compatible integrated CNS tumour diagnostics (**MGMT* promoter methylation test recommended for high grade gliomas)

Therapeutic options for primary brain tumour patients remain limited, particularly after radiotherapy and chemotherapy have failed. NGS panels are increasingly used to identify molecular alterations for targeted therapy, especially when first-line treatments are exhausted. However, the clinical relevance of these alterations varies, from established therapeutic efficacy to hypothetical targets based on preclinical evidence. In order to reduce redundant testing, additional workload and financial burden, molecular testing in gliomas, neuronal or glioneuronal tumours is recommended by the EANO guidelines for recurrent or resistant tumours, rare tumour types or types with no standard care protocols, high mutational burden (TMB) or mismatch repair deficiency (MMR), and for clinical trial participation[61].

In mid-2015, as the WHO CNS tumour classification update was being prepared, DNA methylation, a previously less emphasised molecular feature, began to show significant potential for classification, though it was not yet central to the process. Methylation based profiling has emerged as a trailblazer in the field of CNS tumour diagnostics. Sturm et. al demonstrated the co-localisation of *IDH1* and *H3F3A* mutations with methylation based unsupervised clusters in diffuse gliomas[13]. Armed with first the 450K and then the 850K EPIC Illumina methylation array, this effort culminated into the first large scale methylation-based classification of CNS tumours published by Capper et. al in 2018[62]. This classifier was trained on a reference set of 2,801 well-annotated tumour samples classified into 91 classes. These classes represented different subtypes of CNS tumours, capturing a wide range of entities, many of which are difficult to distinguish using traditional methods. These classes were primarily histology-driven and all samples had established diagnoses. This tool called the MNP or Molecular Neuropathology classifier has been made freely available on the website

www.molecularneuropathology.org. As of 29th September 2024, the website had 158,076 uploads by 1800 users from 500 different institutes globally. Methylation profiles from new cases are constantly added to a centralised database which are then subjected to unsupervised clustering like t-SNE or UMAP analysis. With the growth of the database, new clusters emerge constantly. Samples in these clusters are queried to identify common histology, mutations using NGS panel sequencing, gene fusions and expression using RNA sequencing and copy number alterations using the methylation array. This is finally correlated with survival and/or drug response leading to the discovery of novel diagnostically relevant classes. The classifier has now evolved to identify 184 methylation subclasses, 143 classes, 75 families, and 34 superfamilies. The methylation classifier was able to classify 12% cases that could not be resolved by conventional histology and has been shown to improve risk stratification of tumours as compared to traditional WHO grading[62-64]. Additionally, methylation profiles provide a more objective and reproducible classification, reducing interobserver variability prevalent in histology. Methylation classification has been accepted and adopted by neuropathology centres globally over the past decade. Multiple studies from over the globe have been published describing independent evaluations of the integration of methylation classification into neuropathology diagnostic workflows[63-69]. This has led to the endorsement of methylation profiling by the 2021 WHO classification and the cIMPACT-NOW consortium[70,71]. For instance, classes like high-grade astrocytoma with piloid features (HGAP), diffuse midline glioma, *H3 K27*-altered and diffuse hemispheric glioma, *H3 G34*-mutant that were identified through their specific methylation profiles have now been added to the WHO classification and their detection is recommended by methylation profiling[72-75]. Remarkably, in the case of meningiomas, the methylation classes are especially prognostic since they are able to distinguish between benign, intermediate and malignant types[76-78]. This tool hence reflects the emerging hallmark of non-mutational epigenetic reprogramming for cancers[53].

CNS tumours are thought to arise from progenitor cells that retain characteristics from early developmental stages. Tumours that emerge from such progenitor cells maintain a methylation "fingerprint" that reflects the cell type they originated from while additionally incorporating tumour specific marks. Thus, methylation profiles serve as a developmental map, reflecting their developmental lineage and allowing us to identify the subtype of the tumour. For example, medulloblastomas, the most common malignant paediatric brain tumours, are divided into several subtypes like WNT, SHH, etc., reflecting the developmental origins of the tumour from different regions of the cerebellum or neural progenitor cells leading

to unique methylation profile[79]. Methylation signatures also serve as markers that reflect the cell of origin of a tumour. This is particularly important in CNS tumours, where certain subtypes mimic specific developmental stages. Gliomas, for instance, often reflect the epigenetic profile of neural progenitors or stem cells[80,81]. These profiles help differentiate between tumour types like astrocytomas or oligodendrogliomas. *IDH1* and *IDH2* mutations play a significant role in altering DNA methylation, which is central to cancer initiation and progression in these tumours[82]. These mutations disrupt the normal DNA and histone demethylation processes, leading to an imbalance in histone methylation, including increases in *H3K36* and *H3K9* methylation[83]. This disruption is also linked to increased CpG island hypermethylation in promoter regions of genes involved in early developmental regulation, similar to the CpG island methylator phenotype (CIMP) seen in some cancers. The leading theory is that these methylation changes are caused by the accumulation of 2-hydroxyglutarate, a by-product of the *IDH* mutations, which interferes with α-ketoglutarate—a critical metabolite for enzymes that regulate DNA and histone demethylation[84]. This interference leads to the build-up of repressive histone marks and promoter DNA hypermethylation, which may contribute to cancer progression.

Tumour heterogeneity in CNS tumours has been shown to be driven by epigenetic changes rather than genetic mutations alone[79,85-88]. These epigenetic modifications, especially in methylation, correlate with developmental timing and tissue-specific patterns, allowing for more precise classification of CNS tumours compared to other tumour types where genetic mutations play a more dominant role[89]. In contrast, tumours in other organs, such as breast, lung or colon cancers, are typically driven by somatic mutations, which can override the epigenetic landscape, making methylation classification less specific for these tumour types[84,90]. In addition, mutations like *IDH, BRAF, NF1, TP53*, etc. and key fusions in genes like *NTRK1-3, ROS1, ALK, MET*, etc are relevant for risk prediction and to guide targeted therapies. Next-generation sequencing (NGS) panels specifically designed to target these mutations and fusions are widely utilised in clinical practice[91]. These panels allow for comprehensive molecular profiling of CNS tumours, enabling the simultaneous detection of key genetic alterations such as SNV/Indels as well as relevant gene fusions. Further, gene expression of certain markers like *EZHIP* is required for diagnosis of *EZHIP* expressing tumours. Thus, cDNA sequencing is also performed in suspicious cases for gene expression and fusion detection[92]. The widespread use of methylation arrays, NGS panels and RNA

sequencing has significantly improved the accuracy of tumour classification, prognosis, and the ability to tailor treatments based on actionable genetic targets[93-96].

Thus, WHO guidelines require identification of a wide range of tumour type specific molecular markers. Today, molecular marker evaluation is incorporated into the diagnostic workflow for all major CNS tumour types.

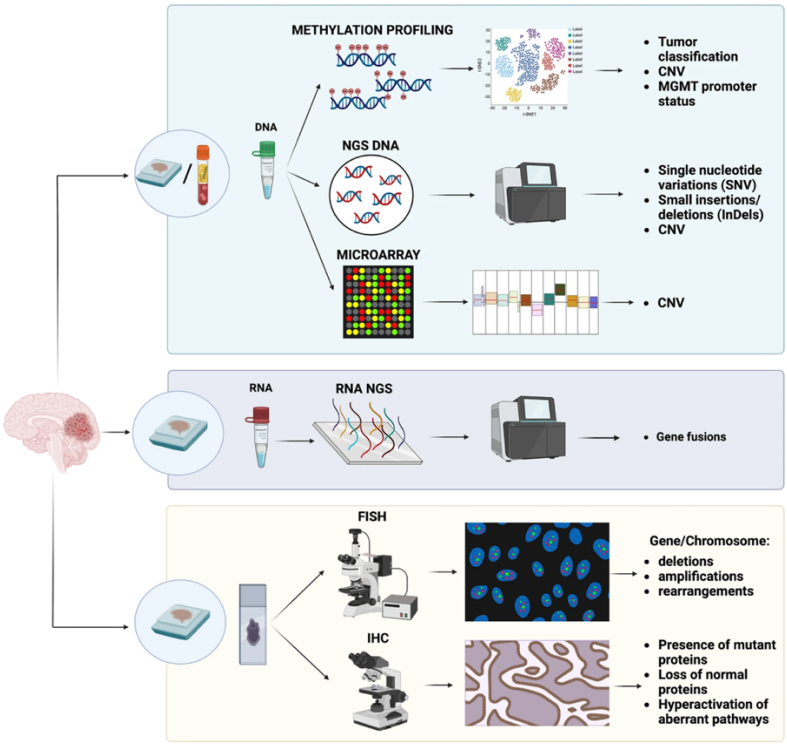


Figure 1-9 Overview of the assays routinely employed in molecular neuropathology diagnostics. (Reprinted from Beretro et. al 2023[97])

As demonstrated in **Figure 1-9,** the methods used for molecular testing include:

NGS DNA sequencing for testing for multiple targets simultaneously to detect gene mutations, small indels and copy number variations with DNA sequencing,

NGS RNA sequencing to detect fusions and gene expression

Methylation arrays to detect copy number variations, *MGMT* promoter status and methylation profiles, including methylation classification.

Immunohistochemistry (IHC) for molecular alterations like *BRAF* p.V600E mutations or *FGFR3* expression, often followed by confirmatory molecular tests. It is not recommended for general use in CNS tumours, except in specific contexts like diagnosing certain fusions.

Fluorescence in-situ hybridisation (FISH) to detect gene amplifications and fusions such as those involving *EGFR*, *PDGFRA*, and *NTRK1-3*. However, it cannot confirm functional gene fusions and is less effective compared to NGS.

The range of equipment required to perform these tests, such as Illumina devices like NovaSeq and the array scanner, can cost in the range of 1-2 million EUR. This has significantly skewed the availability of such testing towards larger centres and developed nations. Furthermore, the library preparation process for these methods is labour-intensive, taking several days and often requiring overnight incubation steps. This demands highly qualified and experienced laboratory personnel, who are often scarce in smaller cities or low-throughput settings. Additionally, the latest versions of the MNP methylation classifier are only compatible with EPIC array data and cannot use data generated by a vast variety of other methylation assays. These limitations have led to criticism of the WHO classification criteria for not being truly universal, as a majority of the global population lacks access to the molecular tests recommended by the WHO.

The primary treatment for central nervous system (CNS) tumours typically involves neurosurgical resection, the extent of which is often determined by the tumour type, as certain types warrant more conservative approaches. For instance, diffuse midline gliomas with the *H3K27* histone mutation are considered incurable, and in such cases, surgery is aimed primarily at obtaining tissue for diagnosis and preserving quality of life, rather than attempting complete tumour removal[98]. Similarly, for medulloblastomas, there is little prognostic difference between near-total and total resection, making maximal resection unnecessary[99]. In contrast, for tumours like posterior fossa ependymoma type A and atypical teratoid rhabdoid tumours, gross total resection is crucial for prognosis, necessitating a more aggressive surgical approach[100-102]. Similarly, in adult CNS tumours, gross total resection has been linked to improved survival in *IDH* wild-type glioblastomas of the receptor tyrosine kinase (RTK) I and RTK II subtypes, but not in the mesenchymal subtype[103]. Similarly, failure to achieve gross total resection in *IDH*-mutant astrocytomas adversely affects overall survival[104]. Therefore, the neurosurgical strategy relies heavily on an accurate and precise diagnosis. Current practice involves preoperative imaging and intraoperative diagnosis through rapid histological assessment of frozen sections. However, this process does not always yield a definitive diagnosis, and post-operative diagnostics may revise the initial findings. Methylation arrays report prognostically relevant copy number profiles and methylation classification using the MNP methylation classifier. However, obtaining results can take several days, making it impractical for intraoperative decision-making[63]. As a result, some patients require a second surgery, while others may have undergone unnecessarily aggressive procedures. The delay in molecular reporting is caused by the sequential nature of conventional methods, for

example, methylation array is ordered after IHC examination, which is often followed by NGS DNA sequencing and also RNA sequencing if needed depending on how the diagnosis shapes as per the previous test results. Additionally, high throughput NGS and arrays are only economically feasible when samples are batched. Thus, in order for a run to be initiated, sufficient samples need to be have accumulated further increasing turnaround time. Even in a high density centre like the Department of Neuropathology at the University Hospital Heidelberg, the average turnaround time for molecular diagnostics is 20 days[95].

In order to improve accessibility and turnaround time of molecular diagnostics for CNS tumours, I developed and validated two tools- Rapid-CNS^2 and MNP-Flex. Rapid-CNS^2 leverages nanopore sequencing technology to target critical genomic regions, delivering both methylation classification and copy number profiling in real-time during surgery—crucial for intraoperative decision-making. Combined with MNP-Flex, a platform-agnostic methylation classifier, this system can offer a broad array of molecular insights required to make a WHO compatible diagnosis.

Chapter 2 Aims of the thesis

To address the challenges outlined in the preceding section, this thesis pursued the following three primary objectives:

- To establish an accessible and swift third-generation sequencing-based pipeline for comprehensive molecular diagnostics of CNS tumours
- To perform an extensive validation of this pipeline within a real-world diagnostic environment
- To develop and comprehensively validate a platform-agnostic methylation classifier that is compatible with the latest versions of the MNP classifier

Chapter 3 Rapid-CNS^2

3.1 Introduction

Nanopore DNA sequencing has emerged as a rapid diagnostic tool, offering several advantages such as low setup costs, compact devices, and real-time data availability[105]. Additionally, nanopore sequencing enables direct measurement of methylated cytosines and significantly reduces sample preparation time. This allows tissue samples to be sequenced early in surgery, potentially providing molecular insights that can inform the surgical approach in real-time. However, a major challenge remains that only sparse methylation profiles can be generated in such a short timeframe, and it is uncertain which CpG sites will be covered. Additionally, shallow WGS results in a 1-2X coverage over the whole genome rendering it inept for mutation or fusion calling. Adding PCR based enrichment faces limitations with long fragments and may erase crucial epigenetic modifications.

Nanopore adaptive sampling is a pioneering technique, rooted in the fundamental principles of real-time sequencing enabled by Oxford Nanopore Technologies (ONT). To address the need for targeted sequencing in nanopore technologies, ONT developed "ReadUntil," a real-time selective sequencing method. This concept leverages the direct interaction with individual nanopores, where the voltage across a pore can be reversed to reject specific reads. As demonstrated in **Figure 3-1**, this provides the potential to either sequence a particular molecule to completion or to eject it mid-sequencing and replace it with another molecule, optimising the sequencing process by selectively reading only the molecules of interest. By quickly rejecting reads from off-target regions, ReadUntil maximises sequencing efficiency, particularly with longer reads, where the loss of sequencing capacity from a single pore can be significant. However, for this selective sequencing to be effective, identification of unwanted molecules must be completed before sequencing is finished. This relies heavily on the speed of sequence identification and the average length of reads. Since read length can exceed 100 kb in some cases, the challenge of the "ReadUntil" approach lies in accurately matching even the shortest fragment of a long read to a reference sequence in real-time. Using the MinION

device, Matthew Loose and colleagues demonstrated the first real-time selective sequencing in 2016[106].

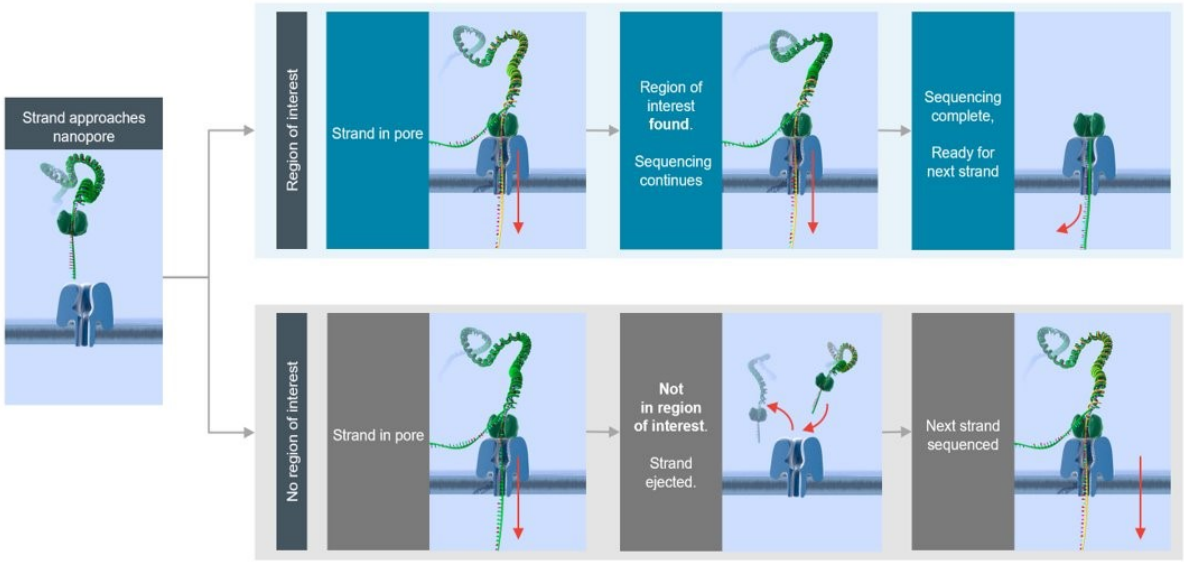


Figure 3-1 Adaptive sampling schematic. (Reprinted from Oxford Nanopore Technologies website https://nanoporetech.com/document/adaptive-sampling, accessed on 4th October 2024)

Building upon the principles of real-time selective sequencing, a method called ReadFish was developed as a robust solution to address some of the limitations of previous approaches like dynamic time warping (DTW) and signal-based methods[107]. DTW was initially used to compare raw nanopore signals to simulated current traces derived from reference sequences, but this method required extensive computational resources, limiting its practicality. Other methods, such as UNCALLED, sought to improve the computational efficiency of raw signal comparison, but they still faced challenges in scalability and resource demands[108]. Direct basecalling of signal fragments was another alternative, offering the advantage of filtering out unwanted reads but without providing true enrichment and still requiring significant CPU power. ReadFish introduced a more efficient approach by focusing on nucleotide sequences rather than raw signals, taking advantage of existing basecalling tools and manageable computational resources. ONT has developed several basecallers, moving from HMMs to neural network-based models that now operate on GPUs. These GPU-based models enable real-time basecalling, allowing data from sequencing flow cells to be processed rapidly enough to match the sequencing speed of the ONT devices, which can sequence up to 2,675 channels simultaneously. By using this real-time basecalling capability, ReadFish integrates with tools like minimap2 to map reads dynamically as they are generated[109]. This eliminates the need to

convert reference genomes into signal space, as required by DTW and other signal-based methods. Readfish follows the logic as demonstrated in **Figure 3-2**.

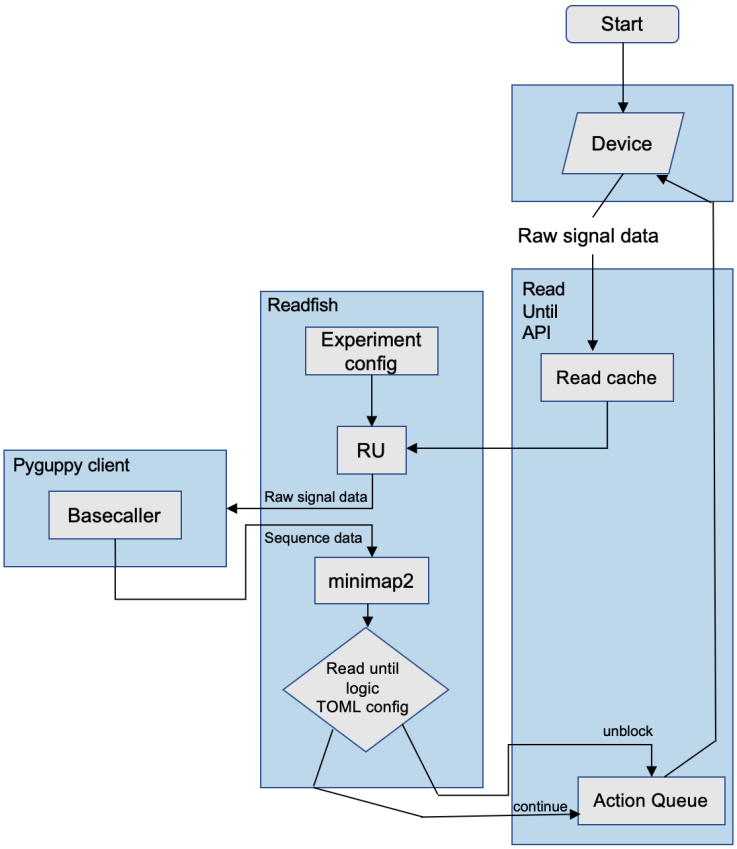


Figure 3-2 Readfish targeting logic flowchart (Adapted with permission from Payne et. al 2021[107])

ReadFish toolkit leverages powerful GPUs, such as the NVIDIA GV100 or RTX 4090, to ensure that even large-scale genomes, such as human chromosomes, can be targeted without computational constraints. In newer updates, ReadFish is able to enrich barcoded samples by barcode balancing further enabling multiplexing of samples[110]. Building on ReadFish, new developments have been made to dynamically update the sequencing focus based on real-time analysis of the reads, enriching specific targets as they are identified during the sequencing run[111].

Multiple studies have reported impressive CNS tumour methylation classification results using nanopore sequencing data. The models trained on the reference data from Capper et. al range from ad-hoc random forests models to neural networks[112-116]. The classifiers have been shown to work on a vast range of tissue types including intraoperatively on frozen sections, FFPE tissue and liquid biopsies. These approaches are restricted to reporting methylation classification only, thus missing out on crucial alterations like mutations, focal copy number alterations and gene fusions. Similarly, due to the random nature of nanopore sequencing, the *MGMT* promoter status, relevant for temozolomide response, is also not reported.

In order to provide a comprehensive range of alterations including methylation classification, mutations, copy number variants (CNV), *MGMT* promoter status and gene fusions, I leveraged adaptive sampling to develop Rapid-CNS^2- a workflow that can report methylation classification and broad CNVs intraoperatively followed by a comprehensive report of alterations on the next day.

In this part of the work, I developed the bioinformatics workflows in bash, Snakemake and Nextflow for targeting and analysis of the data. I utilised community-developed tools for SNV, CNV, and SV calling and annotation. To address gaps in tools, I developed methods for methylation classification and *MGMT* promoter methylation status. I also optimised speed of analysis using multiprocessing and GPU-based methods where applicable. I validated the results using data from matched samples generated using conventional methods like NGS panel sequencing and methylation array.

3.2 Methods

3.2.1 DNA extraction and library preparation

DNA extraction and library preparation were performed by technicians, MD students, masters students and working students at the Department of Neuropathology, University Hospital Heidelberg. *Helin Dogan optimised the protocol using R9 flowcells as follows, quoted from Patel et. al 2024[117]:*

“DNA was extracted using the Maxwell® RSC Blood DNA Kit (Promega, #AS1400) following manufacturer's instructions. In summary, 40x10 µm of fresh frozen tumour tissue were incubated with 300µl of Lysis Buffer and 30 µl of Proteinase K, at 56°C overnight, with continuous agitation at 550 rpm. The following day, samples were transferred into Well 1 of a Maxwell cartridge. DNA extraction was performed using the recommended protocol on the device. For sequencing performed in Heidelberg, two protocols were followed depending on flowcells used. For R9 flowcells, we used the previously described protocol[118]. In summary: DNA concentrations were quantified with the Invitrogen Qubit DNA BR Assay Kit (Q32851, Thermo Fisher Scientific) using a FLUOStar Omega microplate reader (BMG Labtech). DNA was then sheared to 9-11 kb using g-TUBEs (Covaris) at 7200 rpm for 120 seconds, and fragment size was assessed with the Agilent 2100 Bioanalyzer using the Agilent DNA 12000

Kit (5067-1508, Agilent Technologies). Sequencing libraries were prepared with the SQK-LSK109 Ligation Sequencing Kit, incorporating modifications including a 30 min end-prep at 20°C and 65°C, followed by AMPure XP bead clean-up, 5 min elution, and a 60 min adapter ligation at room temperature. The ligation mix was cleaned with AMPure XP beads (0.4x) and eluted in 31 µl using the Long Fragment Buffer. Library concentration was determined with the Invitrogen Qubit DNA HS Assay Kit (Q32851, Thermo Fisher Scientific) on a Quantus fluorometer (Promega). Libraries (500-600 ng) were loaded onto FLO-MIN106 R9.4.1 flow cells with at least 1100 pores available, and sequencing was performed on MinION and GridION platforms, with flow cells flushed twice per sample using the Flow Cell Wash Kit (EXP-WSH003)."

The following protocol describes the protocol for R10 flowcells as optimised by Pauline Göller and Michelle Brehm quoted from Patel et. al[117]:

"2.5 µg of extracted DNA was sheared to 10kb fragments in 60 µl nuclease-free water using Covaris g-TubeTM (Covaris, #520079) following manufactures instructions. Sequencing library was prepared using the ligation sequencing kit (Oxford Nanopore Technologies, SQK-LSK114) and the NEBNext® Companion Module (New England Biolabs, E7180S) with only minor adjustments to the original protocol SQK-LSK114. In brief, DNA repair and end-prep was carried out starting with 58 µl of sheared DNA as input. The ratio of Ampure Beads for bead cleanup was adjusted to the volume of the sheared DNA as proposed by Kolmogorov et. al 2023[119]. Adapter ligation and bead clean-up was performed using the short fragment buffer. DNA was eluted in 15 µl for a MinION sequencing run, and 25 µl for a PromethION sequencing run. If sequencing was performed with a MinION (FLO-MIN114, R10) flow cell, flow cells were primed using the BSA supplement and sequenced using a GridION (Device and software). If sequencing was performed with a PromethION (FLO-PRO114M, R10) flow cell, flow cells were primed without BSA supplement and sequenced on the P2 solo (Device and software). Sequencing on both devices was performed with 600-700 ng of DNA library."

3.2.2 Adaptive sampling

I initially tested Panel A on 47 samples that constituted regions from the neuropathology gene panel and 10,000 CpG sites used for classification by MNP classifier (available on GitHub https://github.com/areebapatel/Rapid-CNS2)[19,120]. I added a 10kb flank to the sites on either side to ensure optimal targeting by ReadFish (155 Mb). Panel B included only the

neuropathology gene panel flanked by 10 kb on either side, with a total targeted size of 15 Mb[120]. All remaining samples were run with Panel B. Both panels were based on the hg19 genome[121]. At the Department of Neuropathology, University Hospital Heidelberg, I ran adaptive sampling in three modes-

1. Using Readfish[122,123] on the MinION
2. Using MinKNOW's in-built adaptive sampling on the GridION
3. Using Readfish on the P2 Solo

I set-up CUDA , Python 3.8 and MinKNOW on a notebook equipped with a NVIDIA RTX 2080 Ti GPU. I installed readfish and read_until_api_v2 in a virtual environment through the https://github.com/LooseLab/readfish git repository (accessed September 2020). I used the fast basecalling mode (config dna_r9.4.1_450bps_fast) from Guppy 4.2.2 to run basecalling for Readfish on the notebook. For the GridION, I used the in-built adaptive sampling option from MinKNOW. I used MinKNOW in the 'offline' mode by disabling pings to the server as required by the firewalls of the University Hospital Heidelberg. I connected the P2 Solo to a local Linux workstation that had 7 NVIDIA RTX 4090 GPUs. I set up Readfish on the workstation through the Git repository (https://github.com/LooseLab/readfish, accessed May 2024). I restricted MinKNOW and Readfish to use one GPU each. If I ran two flowcells in parallel, I assigned one GPU to the Readfish process of each. I ran readfish using the readfish targets command.

3.2.3 Dataset

My dataset included 252 frozen samples- 112 archival and 252 diagnostic samples including 5 intraoperative from Department of Neuropathology, University Hospital Heidelberg. I included 13 samples sequenced intraoperatively at the lab of Matthew Loose at University of Nottingham for the intraoperative analysis part only.

3.2.4 Bioinformatics analysis pipeline

I curated and constantly developed an end-to-end bash pipeline to analyse the data. The pipeline had two major versions- v1 and v2. v1 was used for samples 1-78 and involved separate basecalling, modified basecalling and alignment on a single GPU. v2 was used for all other samples and was capable of simultaneous basecalling, alignment and modified

basecalling by harnessing multi-GPU potential. Kirsten Göbel adapted it to analyse data locally on GPU workstations in the network of the Department of Neuropathology at the University Hospital Heidelberg. I analysed samples 1 – 139 on the ODCF cluster of DKFZ, and Kirsten Göbel analysed samples 140 – 252 using local GPU workstations. Owing to rapid turnover in nanopore sequencing technology and the tools to analyse the data, the pipeline underwent multiple updates in terms of versions and specific tools used. Barring logistical limitations, I employed the latest stable versions of tools available at the time of analysis of the respective samples. I developed bash (v1 and v2), Snakemake (v1) and Nextflow (v2) pipelines to ensure flexibility and adaptability across environments. Pipelines are available on Github as https://github.com/areebapatel/Rapid-CNS2_nf and https://github.com/areebapatel/Rapid-CNS2_sh. I used the NVIDIA Clara Parabricks docker container to run accelerate tools as indicated below. The basic pipeline involved: basecalling, alignment, QC, coverage analysis, methylation calling, methylation value extraction, SNV/CNV/SV calling, variant annotation and filtering, methylation classification, *MGMT* promoter methylation analysis and report generation.

Basecalling: I deployed basecalling on the LSF cluster of the ODCF. For samples 1-78 using v1, I used the guppy basecaller binary in a bash script with a single GPU. For samples 79-145 using v2, I deployed basecalling in multi-GPU mode on the LSF cluster. I used basecaller versions and models as indicated in **Table 3-1**.

This extract is taken from Patel et. al 2024:

“I performed basecalling in a basecall server-supervisor mode for ONT’s proprietary software guppy or Dorado (https://github.com/nanoporetech/dorado). For the multi-GPU mode, I used 15 basecall clients for a 3 GPU setting with available NVIDIA GPU models (RTX 2080 Ti, A100, V100). For a single GPU, I ran guppy_basecall_supervisor or subsequently ont_basecall_supervisor on with 5 clients. For local deployment at the Department of Neuropathology, Kirsten Göbel used a single NVIDIA RTX 3090 Ti GPU powered local workstation or 2 GPUs on the multi-GPU workstation with a Dockerised pipeline.”

Table 3-1 Basecaller versions and models

Samples	***Basecaller version***	***Basecaller model***
1 – 60	Guppy v4.4.1	High accuracy with 5mC
61 – 78	Guppy v5.0.1	High accuracy with 5mC

79 – 139	Guppy v6.1.7	Super accuracy with 5mC
140 – 145	Guppy v6.4.6	Super accuracy with 5mC
146 – 252	Dorado basecall server 7.1.4	Super accuracy with 5mC

For samples using v1, I performed adapter trimming using Porechop, alignment to the hg19 genome using minimap2 v2.18 followed by samtools sorting and indexing[121,124-126]. For all following samples, simultaneous basecalling and alignment was possible. Thus, I merged the bam files into a single bam file and indexed it using samtools[125].

Methylation calling: In v1, I performed methylation calling for samples 1-78 using megalodon v2.3.3 with a guppy backend (https://github.com/nanoporetech/megalodon). For all following samples using v2, I used the built-in capabilities of the basecaller to directly output methylation tagged bam files. I used the basecalling models as indicated in **Table 3-1**. I extracted methylation values using modbam2bed (https://github.com/epi2me-labs/modbam2bed) with the `--cpg` parameter. I performed liftover of the methylation bed files to the hg38 genome using the liftOver tool[127].

SNV calling and annotation

I performed SNV calling using the latest available version of DeepVariant[128] on the reads mapping to the targeted regions. For v2 run on the ODCF cluster, I used the Parabricks accelerated version of PEPPER-Margin-DeepVariant on a GPU[129,130]. The subset bam file was generated using the `bedtools intersect` function[131]. I annotated SNVs using ANNOVAR[132]. Filtering for clinical relevance was based on the 1000 Genomes (Aug 2015) frequencies and COSMIC 68 database[133,134]. For pathognomonic alterations in *IDH1/2*, *TERTp*, *BRAF* V600E, *H3F3A* and *H3K27M,* I additionally ran `bcftools mpileup` over the relevant regions[125,135].

CNV calling

This extract was adapted from Patel et. al 2024:

"I called copy number variations on the entire bam file with bin sizes of 1kb, 10kb and 100kb using default parameters for cnvpytor[136]. I plotted the copy number profiles generated using a 100kb bin size. Copy number status of relevant genes was reported using a custom python script. The script parses the pytor file obtained as output of cnvpytor. If the complete gene was covered by the bin, the copy number status of the bin was assigned to the gene."

Methylation classification

I developed a custom random forest classifier to analyse DNA methylation profiles of central nervous system tumours derived from nanopore sequencing. The model was created and run in R 4.2.0, using the publicly available 450k methylation array reference dataset from the MNP methylation classifier version 11 (GSE90496), as pre-processed in Capper et al. 2018[19]. For each nanopore sample, I selected methylation calls that overlapped with the top 100,000 probes (ranked by mean decrease in accuracy) from the MNP classifier. I then applied variance filtering to these probes, narrowing it down to the 10,000 most variable ones. Using these probes, I trained a random forest model with 20,000 trees, using the `ranger` package[137]. To improve precision of the model, I recalibrated it by training one-vs-all generalised linear models for each class. These models produced a confidence score for each prediction. I determined the methylation families by grouping similar methylation classes from the reference set.

For the samples with corresponding tissue analysed using the EPIC array v1, I validated the results based on the MNP methylation classifier v11b4 predictions. However, samples processed with the EPIC v2 chip lacked v11b4 predictions, so I inferred annotations based on the MNP v12.8 classification. This involved: a) assigning a specific class if it matched one class in v12.8, b) assigning a methylation family if multiple subclasses were present, or c) labelling the sample as 'Not in classifier' if no corresponding class was present in MNP v11b4.

***MGMT* promoter methylation**

I inferred ground truth for *MGMT* promoter methylation status from EPIC array analysis. I split 59 samples (47 Panel A and 12 Panel B samples) into 70% training and 30% validation data. I subjected each of the 212 CpG sites in the *MGMT* promoter region to a Student's t-test to assess their predictive value. I selected 137 sites with p-value <0.01. I trained a logistic regression based binomial classifier on the average of these 137 sites. I subset the hg38 bedmethyl file for each sample to `chr10:129466536-129467536`. I selected the aforementioned 137 sites and calculated the average. I ran the *MGMT* prediction model on the results.

SV calling

I performed SV calling using svim and Sniffles[138,139]. I ran Sniffles in non-germline mode with a minimum support of 2X for reporting variants. I annotated the variants using AnnotSV[140]. For fusion calling, I manually queried the VCF for genes relevant to the predicted methylation class or suspected diagnosis. I further visualised the results in IGV[141].

Report generation

I generated reports using R packages `rmarkdown`, `kableExtra`, `knitr`. The report generation scripts used outputs from the coverage calculation, filtered SNV calls, copy number variation profile, methylation classification and *MGMT* promoter prediction. I generated reports in the PDF and HTML formats.

3.2.5 Integrated diagnosis

Pathologists provided integrated diagnoses by considering histology, clinical data, and molecular findings from the Rapid-CNS^2 workflow. For conventional analyses, the integrated diagnoses were based on histology, clinical data, and results from the MNP v12.8 classifier, along with DNA panel sequencing and/or RNA sequencing, depending on availability[19,120,142,143].

3.2.6 NGS and EPIC sequencing and analysis

NGS and EPIC array analysis were performed as described previously in a routine setting at the Department of Neuropathology, University Hospital Heidelberg[19,120,144].

3.2.7 Intraoperative sequencing simulation

This extract was adapted from Patel et. al 2024:

"For 39 Rapid-CNS^2 samples sequenced in Heidelberg, I conducted simultaneous basecalling and alignment to the hg19 genome using guppy 6.4.6 with the super accuracy configuration and 5mC modification detection[121]. Sequencing summary files aided in extracting cumulative reads at time intervals (5 min, 10 min, …. 1440 min). I extracted methylation values using modbam2bed (https://github.com/epi2me-labs/modbam2bed), performed liftover to the hg38 genome and applied the ad-hoc Rapid-CNS^2 classifier[118,127]. Simultaneously, I conducted CNV calling using QDNAseq with a bin size of 1 Mb[145]."

3.2.8 Intraoperative sequencing protocol

Intraoperative sequencing was performed on 5 samples by Jochen Meyer at the Department of Neuropathology, University Hospital Heidelberg and for 13 samples by Simon Deacon at University of Nottingham. For the protocol, we require minimum 5 mg tissue. The lab of Matt Loose in Nottingham developed the initial protocol. The samples were prepared with the ONT ultra-long kit using an adjusted protocol (SQK-ULK114) (https://protocols.io/view/intra-operative-nanopore-sequencing-to-classify-br-c65qzg5w). For intraoperative samples sequenced in Heidelberg, Jochen modified an adjusted protocol using the ONT ultra-long kit to perform protein cracking using Maxwell-DNA-extraction by a PreCellys-cell/tissue-homogenization-device instead of shearing by needle. The samples in Nottingham were analysed in real-time using the ROBIN pipeline and a custom pipeline at Heidelberg[146,147]. Briefly, the pipelines basecall and align the FAST5/POD5 files using dorado with the high accuracy model for 5mC detection as soon they are written to the output folder. Methylation values were extracted using modbam2bed at Heidelberg and modkit at Nottingham. Both ran the Rapid-CNS[2] methylation classifier on the files upon generation. At Heidelberg, we ran QDNAseq for copy number variant calling on bam files at 5-minute intervals.[145] The bedmethyl files for every 5 minute intervals from Nottingham were sent to me through OneDrive. I re-ran the classifier for the files and plotted the results using `ggplot2`.

3.3 Results

3.3.1 Adaptive sampling vs whole genome sequencing

To assess the efficiency of adaptive sampling, I compared 5 whole genome sequencing libraries to 5 adaptive sampling libraries sequenced on a GridION. I found a considerably higher number of reads in the adaptive sampling libraries. WGS libraries showed an increased N50 and median read lengths compared to the adaptive sampling ones (**Figure 3-3** a) to c)). Adaptive sampling involves a real-time rejection of reads, ideally within the first 180 bp. Panel B covers 1.5% of the genome. Assuming relatively uniform distribution of reads over the genome, it follows that most reads would be rejected with read lengths of ~200 bp. Thus, it would be expected that adaptive sampling libraries accumulate higher number of reads coupled with lower N50 and median read length as compared to WGS libraries. **Figure 3-3** d)

displays a significant increase in coverage over the on-target region for adaptive sampling libraries compared to WGS, confirming the efficacy of real-time targeting using adaptive sampling. Despite exclusion of CpG sites from panel B, a greater number of CpG sites overlapping the 450k array sites were detected by adaptive sampling libraries potentially owing to the greater uniformity of rejected reads. CNV profiles of adaptive sampling libraries also displayed better resolution than those from WGS libraries for the same bin sizes. This could also be attributed to the contribution of the short rejected reads that covered a larger scope of the genome in a shorter time as opposed to the much longer reads in WGS. For example, in the time taken to sequence a 10,000 bp read covering a single region of the genome, an adaptive sampling library could cover around 50 (10,000 ÷ 200) different genomic regions.

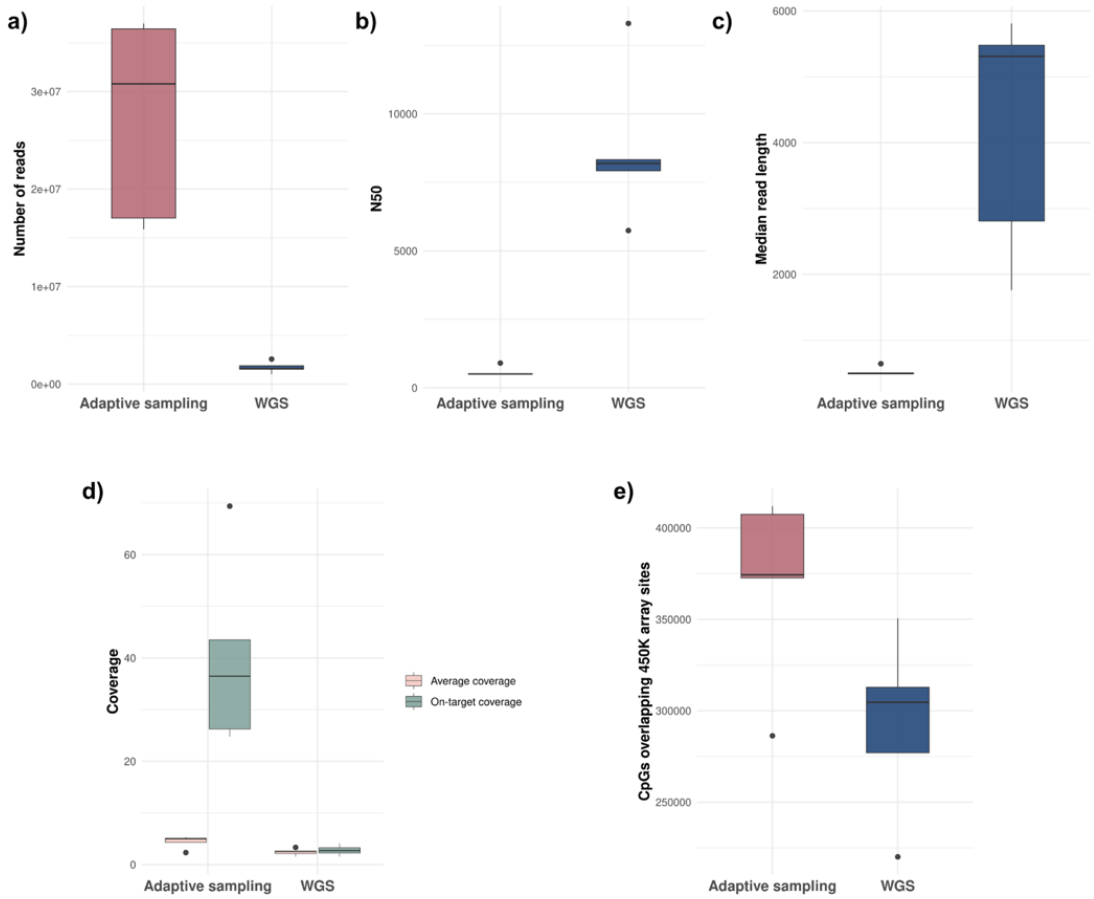


Figure 3-3 Comparison of whole genome sequencing and adaptive sampling libraries

3.3.2 Workflow establishment

Helin Dogan and I established the wet-lab and computational workflow for Rapid-CNS[2] respectively. I set up the adaptive sampling tool Readfish on a consumer grade notebook, while Helin Dogan optimised incubation time and other wet-lab parameters for efficient targeting and output[118].

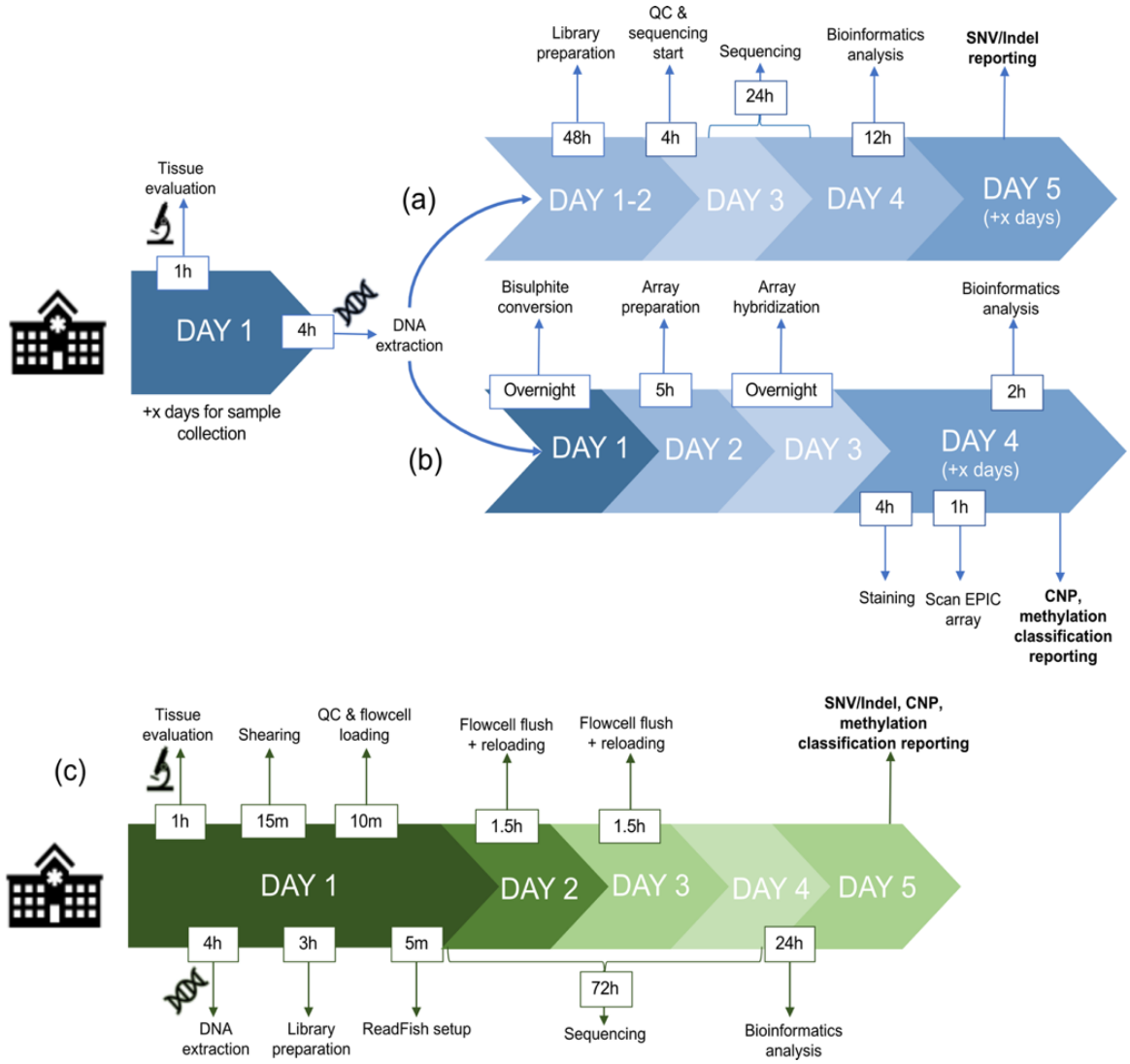


Figure 3-4 Comparison of timelines for a complete molecular workup with conventional methods -NGS panel seq (a) and EPIC array (b) to Rapid-CNS^2 v1 (c). (Reprinted from Patel et. al 2022[148])

Conventional molecular analysis involved NGS panel sequencing and EPIC array analyses. **Figure 3-4** demonstrates the timeline comparison for Rapid-CNS^2 vs conventional analysis. The Rapid-CNS^2 workflow reduces the library preparation time to 7h as compared to over 48h for conventional analyses. Conventional analyses suffer from a need to batch multiple samples for cost effectiveness, exacerbating the delay in turnaround time. Rapid-CNS^2 improves upon that by allowing single sample processing, making it possible to start sequencing upon sample receipt.

We first sequenced 45 archival frozen glioma samples for 72 h each to examine the feasibility for identifying pathognomonic alterations using a handheld MinION (**Figure 3-5**). Targets

(Panel A) included the regions of the neuropathology panel and 10,000 CpG sites inferred from the MNP classifier (155 Mb)[19,120]. I compared reported molecular alterations to conventional data from corresponding FFPE tissue samples. I could reliably detect *IDH1* R132H/S mutations in 12/13 samples and *TERT* promoter mutations in 23/25 samples with mutations detected in NGS panel seq. Pathognomonic copy number alterations including focal alterations like *EGFR* and *CDKN2A/B* were correctly identified. *MGMT* promoter methylation status was accurate in 45/45 samples. Methylation families were accurately identified in all 45 samples and methylation classes were also correct in 37 out of 45 cases.

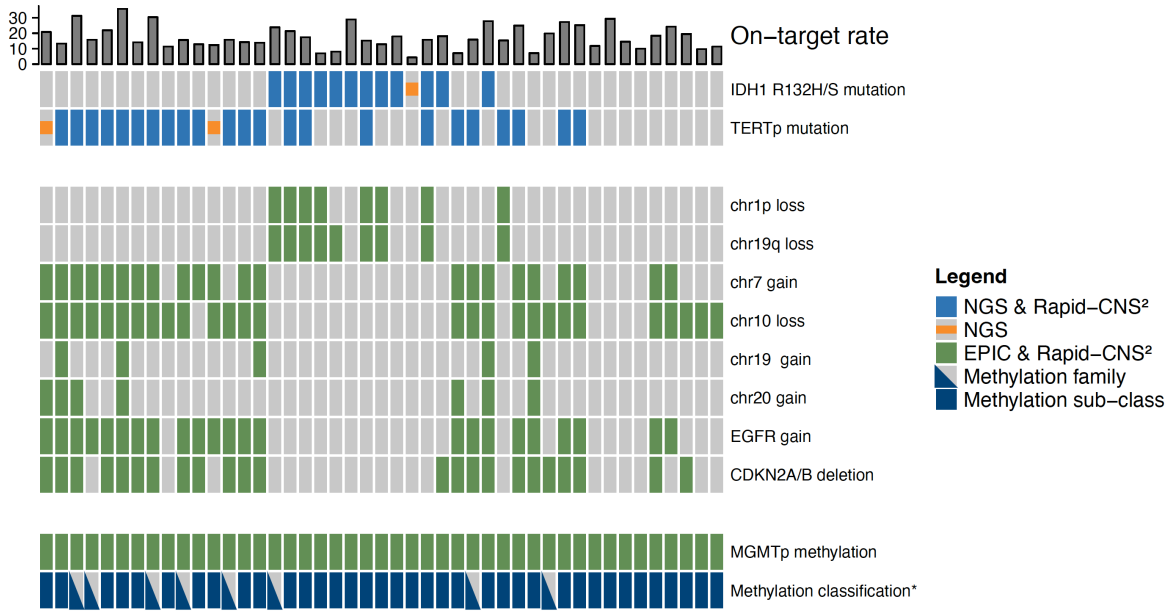


Figure 3-5 Concordance of pathognomonic alterations for glioma samples targeting Panel A

To further improve coverage, we targeted Panel B which only included regions from the NGS gene panel[120]. We then sequenced 8 archival frozen glioma samples that targeted Panel B on a GridION. Similar to previous results, I found complete concordance for pathognomonic SNVs, CNVs, *MGMTp* status and methylation family classification (**Figure 3-6**).

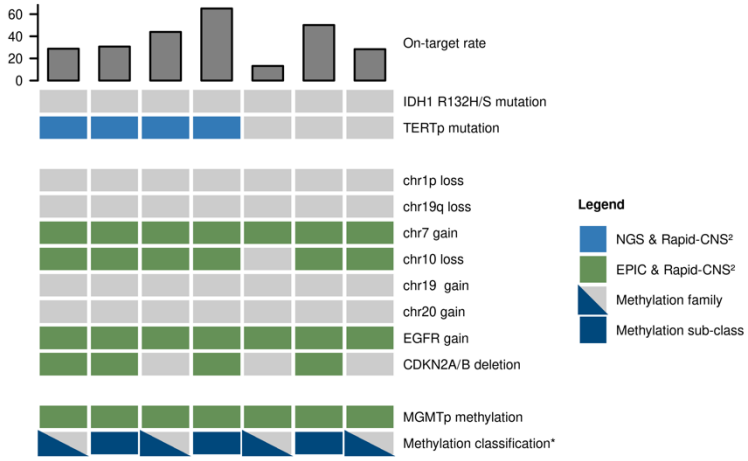


Figure 3-6 Concordance of pathognomonic alterations for glioma samples targeting Panel B

3.3.3 Panel curation

As shown in **Figure 3-7,** libraries sequenced using the smaller panel B (15 Mb) achieved a higher on-target rate compared to those sequenced using the larger panel A (155 Mb). Despite removal of CpG site-specific target regions in B, the number of detected CpG sites overlapping the Illumina 450K methylation array sites remained consistent across panels. CNV profiles also maintained consistent resolution. This could largely be attributed to off-target reads. Thus, B was designated as the default panel for subsequent runs.

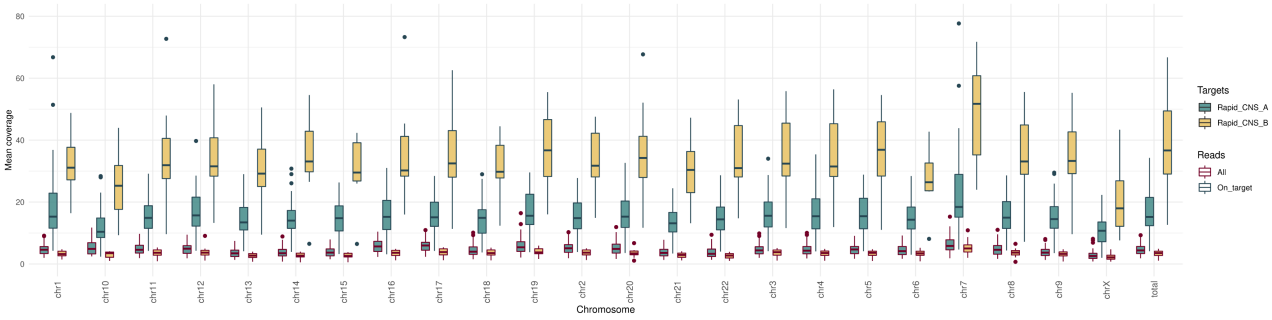


Figure 3-7 Mean on-target coverage for samples targeting Panel A vs Panel B. (Reprinted from Patel et. al 2022[148])

3.3.4 Sequencing time optimisation

I split five libraries run using Panel B into two each, containing reads generated in the first 24h and those generated in the subsequent 48h after flushing and reloading respectively. Since flushing the flowcell depletes all previously loaded sample, I considered data generated after reloading the flowcell to be equivalent to loading a new library on a previously used flowcell. Since the same sample was loaded again, it avoided any sample or flowcell-related bias.

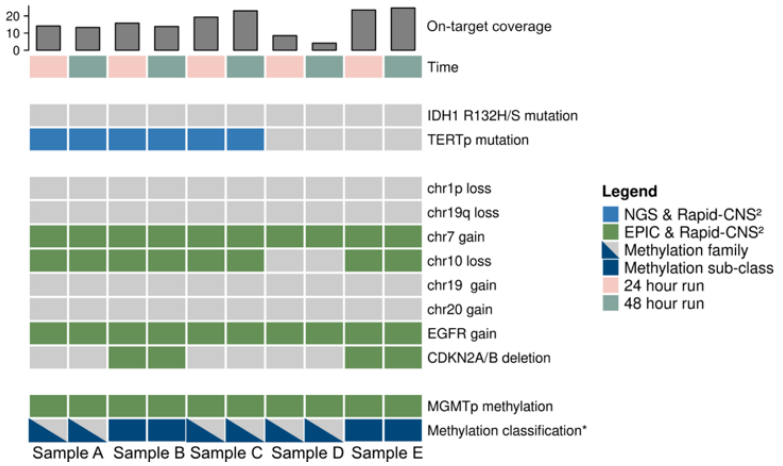


Figure 3-8 Oncoprint showing concordance for samples sequenced for 24h and after flowcell washing and reloading for 48h

Each split library contained > 5 million reads. While there was no clear trend observed for the number of reads generated in the first 24h vs next 48h, mean on-target coverage for all libraries was 10-15X, similar to that observed with Panel A libraries sequenced for 72h. As shown in **Figure 3-8**, complete concordance was reported for all pathognomonic alterations (*IDH1, TERTp, MGMT* promoter methylation and copy number alterations). Methylation families were correctly identified in all split libraries, 4 of which also reported the accurate methylation sub-class as identified by the corresponding EPIC array-based classification (same as their corresponding 72h runs).

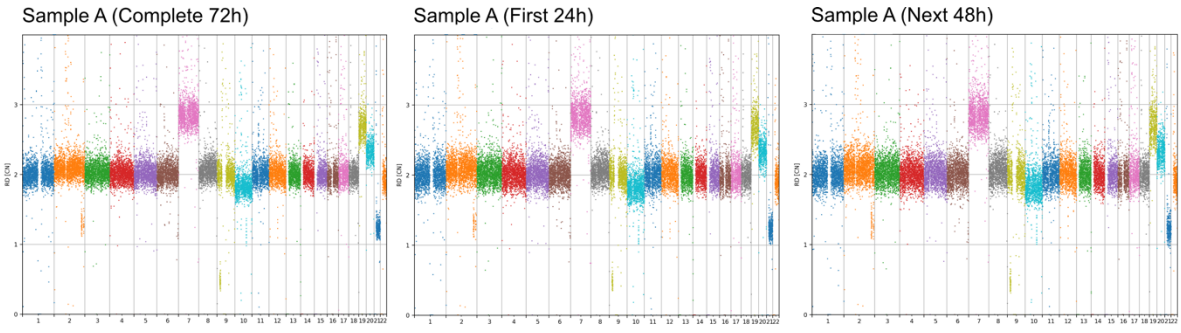


Figure 3-9 Copy number profiles at varying sequencing times.

Copy number profiles were identical for the complete 72h run, first 24h run and next 48h run for all split libraries (**Figure 3-9**). For a 24h run, sequencing with the shorter panel (Panel B) resulted in mean on-target coverage equivalent to a 72h run with the longer panel (Panel A). This demonstrated that sequencing time can be reduced proportionally by decreasing target sizes. As flowcell quality was maintained, it was decided to sequence all samples after Panel B for 24h or after washing and reloading for 48h.

3.3.5 Sample overview

I collected a dataset of 301 samples to develop and validate the pipeline. 252 samples were sequenced at Department of Neuropathology, University Hospital Heidelberg and 49 samples were sequenced at DeepSeq, University of Nottingham. For the sake of completeness of analysis and since they were locally processed, I only refer to the 252 samples sequenced in Heidelberg in this chapter except for when specifically mentioned. Of these, 112 samples were retrospectively sequenced, while 140 samples were prospectively sequenced in a 'real-life' diagnostic setting. Among the scope of tumour types included in the dataset, the cohort included particularly 53 molecular low-grade tumours (e.g. pilocytic astrocytoma, ganglioglioma, dysembryoblastic neuroepithelial tumour, and CNS WHO grade 2 diffuse glioma), 15 recurrent tumours, 10 samples with infiltration zones of diffuse glioma, and 5 samples the size of a small biopsy (~1.5mm diameter). Of these, we sequenced 5 samples in an intraoperative setting where the samples underwent library preparation and sequencing as soon as they were sent to the Department of Neuropathology for frozen section inspection. Kirsten Göbel ran live analyses including methylation classification and copy number profiling on these samples. To provide a better understanding of the intraoperative sequencing and results, I also added 13 samples sequenced intraoperatively at the University of Nottingham.

3.3.6 Coverage

Adaptive sampling resulted in a noticeable improvement in on-target coverage for all libraries. I observed a significant difference in coverage depending on the device used, in-line with the specifications. P2 Solo provided the highest on-target coverage, achieving an average of 44.8X after 24 hours of sequencing. In comparison, the GridION device showed varying coverage levels, with an average of 16.5X across all runs. Specifically, 72-hour runs on the GridION yielded 24.4X coverage, while shorter runs (24 and 48 hours) averaged 12.56X. The MinION, though the smallest device, performed well with an average on-target coverage of 17.5X during 72-hour runs. It should be noted that the samples sequenced on the MinION targeted Panel A which is 10 times the size of Panel B employed by the other libraries. These results demonstrate that while all devices benefited from adaptive sampling, the P2 Solo was the most effective in reaching higher coverage in a shorter period, with the GridION and MinION offering reasonable alternatives with lower, but still reasonable coverage.

3.3.7 SNV calling

I compared SNVs from 103 samples to their matched NGS libraries. I compared all variants called using DeepVariant for the Rapid-CNS[2] samples and mpileup for their corresponding NGS libraries. I observed 91.7% recovery of SNVs called in NGS data (**Figure 3-10**).

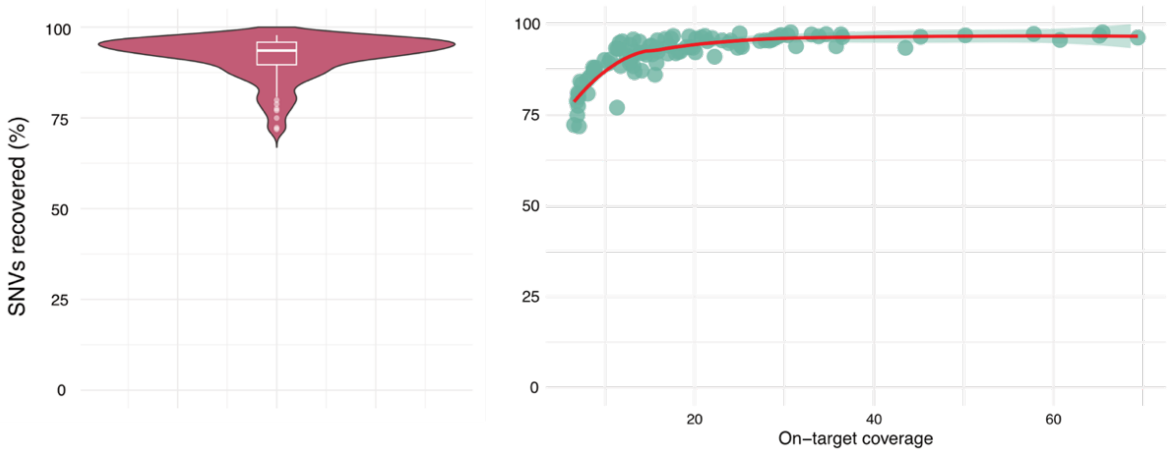


Figure 3-10 SNV concordance and its relationship with on-target coverage. (Reprinted from Patel et. al 2024[117])

Since targeting by adaptive sampling was primarily meant to improve variant calling over clinically relevant regions, I delved into the relationship between SNV recovery and on-target coverage. I found that proportion of SNVs recovered increases with on-target coverage A minimum on-target coverage of 10X was necessary to capture 90% of the variants (**Figure 3-10**).

Minor differences between sequencing platforms are typically expected and unavoidable due to the inherent distinctions between nanopore and Illumina technologies. However, these differences can pose challenges when they involve clinically relevant alterations. Despite technical variability across sequencing platforms, pathognomonic and canonical mutations that align with the integrated diagnosis, as well as those confirmed by direct sequencing or by mutation-specific antibodies at the protein level, can be regarded as the biological "gold standard" or "ground truth." Hence, I investigated *IDH1/2* and *BRAF* mutations and found correct calling in 44/45 and 1/1 samples respectively with matched NGS data and endorsed by direct sequencing and/or immunohistochemistry, with no false positives (97.9% sensitivity, 100% specificity). For samples with methylation class predictions suggestive of an *IDH1/2* mutation, I consulted the mpileup variant calls if the variant was not called by DeepVariant.

TERT promoter mutations are notoriously hard to amplify for NGS and subsequently detect owing to their presence in a GC-rich region[149]. The resulting poor coverage makes variant calls in the region susceptible to false negatives and false positives. Previous work found over 65 % false negatives in variant calling with targeted NGS over Sanger sequencing which could only be resolved by looking at raw reads in IGV[149]. In line with previous findings, coverage over the *TERTp* region for our NGS panel sequencing was ~10% of the mean on-target coverage for the library. Thus, establishment of ground truth in this region is a matter of some debate. However, considering NGS calls as true mutations, I found that they were concordant in 48 out of 54 (88.9%) cases (A.1.1).

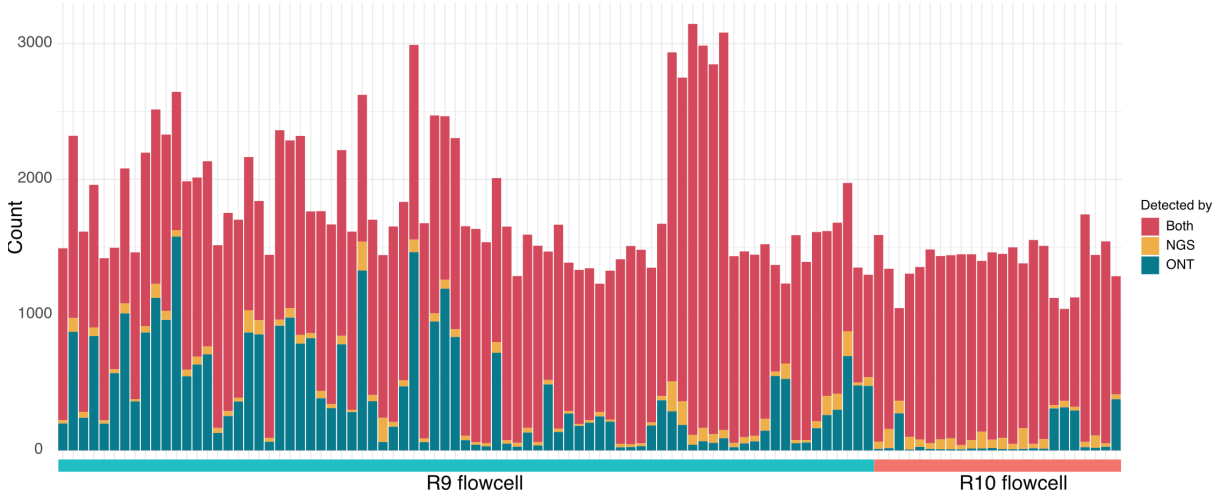


Figure 3-11 Number of shared and unique mutations per sample. (Reprinted from Patel et. al 2024[117])

To examine the complete set of variants, I probed the variants detected by NGS panel sequencing and Rapid-CNS^2 from each sample with available corresponding data (**Figure 3-11**), considering NGS variant calls as the reference or "true" calls. I found 68,941 true positives, 35,370 false positives, 6170 false negatives. However, previous studies have demonstrated substantial variability in the performance of different variant callers when applied to the same dataset[150]. In this analysis, the sequencing technologies, tissue types (FFPE for NGS panel sequencing), and variant calling algorithms used are distinct, making the assumption that NGS calls represent absolute ground truth not entirely accurate. Therefore, the discrepancies observed in variant calls between the two methods are consistent with findings from literature.

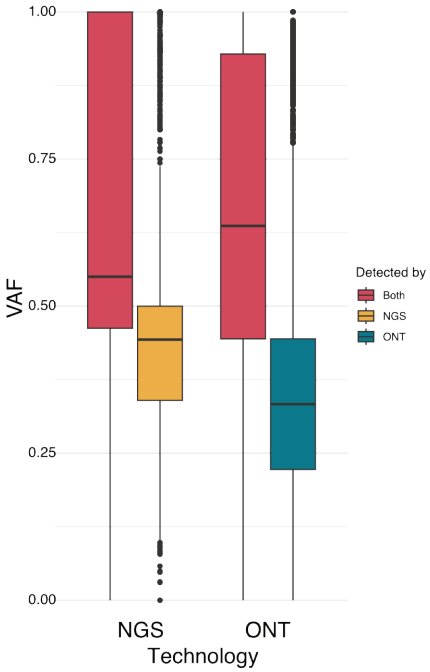


Figure 3-12 Variant allele frequency comparison of detected mutations. (Reprinted from Patel et. al 2024[117])

Additionally, I queried the variant allele frequencies (VAF) of shared and exclusive variants in NGS and ONT (Rapid-CNS^2) libraries. The results showed that the VAFs of variants detected by both technologies were significantly higher than those found exclusively by either method (**Figure 3-12**). This indicates that the variants missed by Rapid-CNS^2 were likely low-frequency mutations, which could represent either false positives or sub-clonal variants. Similarly, false positives identified by Rapid-CNS^2 were associated with low VAFs, suggesting that these could potentially be filtered out by applying a VAF threshold.

As evident from **Figure 3-11**, the number of false positives decreased considerably with R10 chemistry. R9 flowcells accumulated and average of 424 false positives per sample, while R10 flowcells improved by 5.5 fold with 77 false positives on average per sample. This aligns with the improved basecalling accuracy claimed by R10 flowcells. As poorer results were obtained from the older R9 flowcells and software versions, which are now obsolete, this clearly demonstrates that advancements in sequencing chemistry and variant calling algorithms play a crucial role in improving the accuracy of SNV calls. Although I evaluated SNV calling across all flowcells, it is important to note that future users will use R10 or higher flowcells, which is guaranteed to yield more reliable results.

3.3.8 *MGMT* promoter methylation analysis

Molecularly, patients with glioblastoma that have hyper-methylated promoter region of the gene encoding O6-methylguanine-DNA methyltransferase (*MGMT*) benefit from alkylating agents like temozolomide as compared to those patients whose tumours lack such *MGMT* promoter methylation[151]. Additionally, extensive resection is more beneficial for patients with *MGMT* methylated glioblastoma[152,153]. Thus, accurate reporting of *MGMT* promoter status is crucial. Conventionally, this is determined using pyrosequencing or Bady's two-site model for the methylation arrays[154]. Due to inconsistencies in the identification of modified sites, the two-

CpG MGMT-STP27 method developed by Bady could not be directly used for assessing the *MGMT* promoter methylation status[2]. I formulated a model that uses the average over multiple predictive sites instead of a site-specific model to deal with the issue of random missingness in high-confidence methylation calls prevalent in nanopore sequencing data. First, I tested a naïve average-based approach. I calculated the mean of methylation values over all reported CpG sites in the UCSC-annotated CpG island. Based on ground truth values for 25 samples, I chose a cut-off of 30%. On further addition of samples, I found that this approach resulted in values being poorly resolved between methylated and unmethylated samples (cut-off ± 5%).

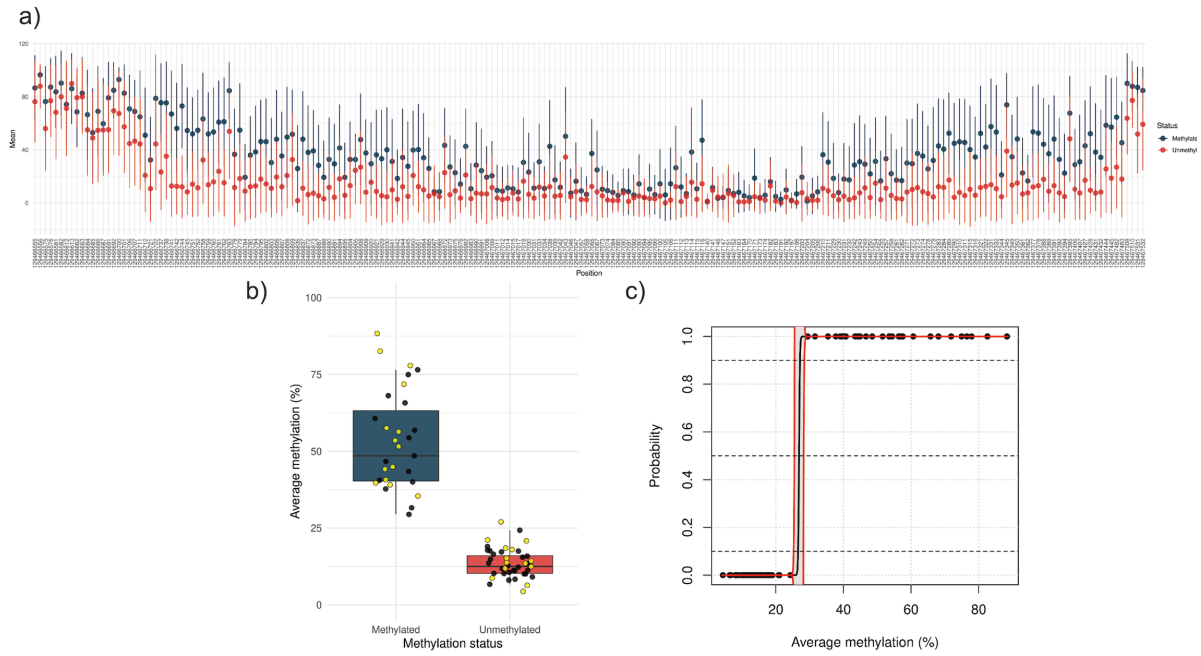


Figure 3-13 *MGMT* promoter methylation status prediction model. (Reprinted from Patel et. al 2022[148])

After comparing the average methylation levels in 59 methylated and unmethylated samples analysed using Rapid-CNS[2] v1, I found that certain sites had poor predictive power (**Figure 3-13** a). On subjecting each site to a Student's t-test, I selected 137 out of 212 sites (p-val < 0.01). When the methylation values were averaged over the 137 selected CpG sites, I observed a clear difference between methylated and unmethylated samples (**Figure 3-13** b). The logistic regression classifier showed in **Figure 3-13** c), which was trained on the training samples and used a 25% cutoff, accurately predicted the methylation status of the test samples. I similarly employed the model for *MGMT* promoter status prediction in Rapid-CNS[2] v2. Methylation calls over the *MGMT* promoter region provided a high resolution of methylation patterns across glioblastoma samples as demonstrated in **Figure 3-14**. The top panel has

aligned reads coloured by sample with CpG sites marked as closed if they are methylated and open if they are unmethylated. The curves on the bottom panel indicate smoothed methylation profiles coloured by sample. As evident from **Figure 3-14**, the profiles show an unambiguous difference between unmethylated and methylated samples.

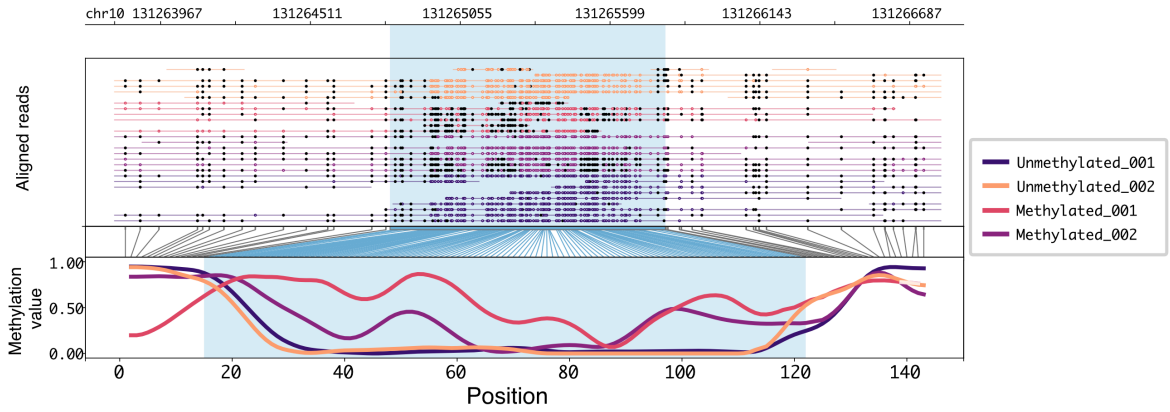


Figure 3-14 Methylation values over the *MGMT* promoter region. (Reprinted from Patel et. al 2024[117])

MGMT predictions remained consistent with the corresponding EPIC array ground truth in 188 out of 207 cases (90.8%) that had matched EPIC array predictions and coverage > 3X over the *MGMT* promoter region (**Figure 3-15**). Three samples with matched EPIC array predictions exhibited low coverage over this region. This discrepancy is consistent with previous reports comparing other *MGMT* methylation assays, such as pyrosequencing versus methylation arrays[154-156]. Notably, discrepancies in *MGMT* methylation predictions are known to occur even between well-established conventional methods, making it challenging to define a definitive ground truth for *MGMT* methylation status[57].

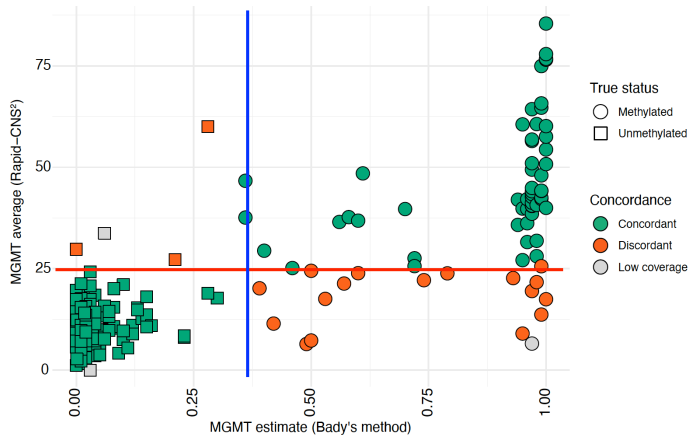


Figure 3-15 *MGMT* promoter methylation concordance

MGMT promoter methylation status is a critical prognostic marker and predictor of response to temozolomide therapy in gliomas, making it essential that this information is accurately conveyed to the treating physician[157-160]. Additionally, *IDH*-wt glioblastomas with methylated *MGMT* promoter have been shown to

benefit from extensive resection in terms of survival[152]. Currently, there is no reliable histopathological feature that would enable a pathologist to determine the methylation status of the *MGMT* promoter. Commonly employed nanopore-based rapid methylation classification methods, such as shallow whole genome sequencing, rely on stochastic, non-uniform coverage of the genome in a short period of time[161-163]. While this approach provides sufficient coverage of CpG sites for training methylation classification models, it does not guarantee coverage of the *MGMT* promoter region. Moreover, methylation classes have not been shown to segregate based on *MGMT* promoter methylation status, meaning that methylation classification cannot serve as a substitute for determining *MGMT* status, unlike the *IDH1* mutation. Therefore, accurate reporting of *MGMT* promoter methylation, alongside methylation classification, SNVs, fusions, and CNVs, enhances the value of Rapid-CNS^2 as a comprehensive diagnostic tool in clinical practice. More importantly, I hypothesise that the improved resolution of the *MGMT* promoter region achieved through Rapid-CNS^2 could serve as an invaluable resource for enhancing patient stratification in predicting temozolomide response.

3.3.9 Methylation classification

I set up the methylation classification scheme using the 59 samples used for workflow establishment. I developed an ad-hoc random forest-based classifier that re-trains on sample specific sites. Similar approaches have been successfully applied elsewhere[161,162,164]. I used the reference dataset from the Capper et. al 2018 paper as reference[19]. Data obtained from a full run usually covered >300,000 probes from the 450K array. Loading the entirety of the overlapping probe set from the training set for re-training warranted considerable memory. I inferred CpG importance from the MNP methylation classifier by selecting sites from the top 100K probes to re-train the ad-hoc classifier. This considerably reduced the time and memory required to perform methylation classification and used a refined feature space. On average, methylation classification (including I/O processes) took 10 minutes with 32 threads. Of the 10,000 probes selected for training in each sample, only ~1400 probes were common in all of the samples. The out-of-the bag error for ad-hoc classifiers for each sample was between 0.18-0.20 (A.1.2). This classifier, referred to as the Rapid-CNS^2 methylation classifier, was applied to all samples of the cohort.

Methylation classification with the integrated Rapid-CNS^2 model covered 91 CNS tumour classes as published in the Capper et. al paper[62]. Of the 228 samples that could be categorised by the conventional methylation classifiers (v11 or v12), 213 (93.4%) were accurately assigned to the correct methylation family, which is generally the most critical diagnostic level.

Since the conventional random forest classifier employs a confidence cut-off of 0.9 for cases considered "classifiable," I derived a corresponding cut-off from the Rapid-CNS^2 data. By filtering samples with a minimum confidence score of 30%, 181 out of 185 (97.8 %) were correctly predicted (**Figure 3-16**). However, to provide a comprehensive assessment, I did not apply this confidence cut-off for the following evaluation of individual Rapid-CNS^2 samples, instead including all cases regardless of score.

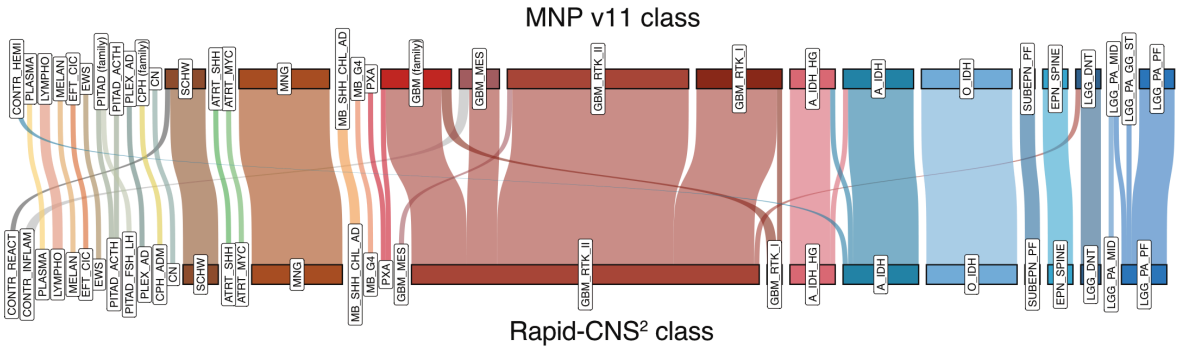


Figure 3-16 Comparison of array-based methylation class predictions to Rapid-CNS^2 predictions (prediction confidence > 30 %)

At the lower hierarchy level of methylation class, 136 out of 189 samples (71.9%) with class-level information available were correctly classified. For instance, 34 cases were concordant at the family level but discordant at the class level within the methylation family of glioblastoma, *IDH* wildtype. Despite having class-level predictions, the WHO classification does not endorse further sub-classification within glioblastoma, *IDH* wildtype, due to the absence of established clinical relevance. Moreover, it has been demonstrated that class-level methylation patterns can vary within different regions of the same tumour sample. This likely accounts for some of the discrepancies observed between frozen and FFPE tissue in certain cases. Therefore, based on current evidence, no biological ground truth for class-level distinctions or their intra-tumoural consistency exists, nor has their clinical significance been established.

Among the 15 samples that did not match their methylation classification, seven were associated with scores indicative of inflammatory or reactive tumour tissue. Of these, four were eventually diagnosed as glioblastoma and received scores corresponding to an "inflammatory glioblastoma microenvironment," which, while not an exact match to the glioblastoma score from the methylation array (which was derived from different tumour regions), still provided sufficient diagnostic clarity.

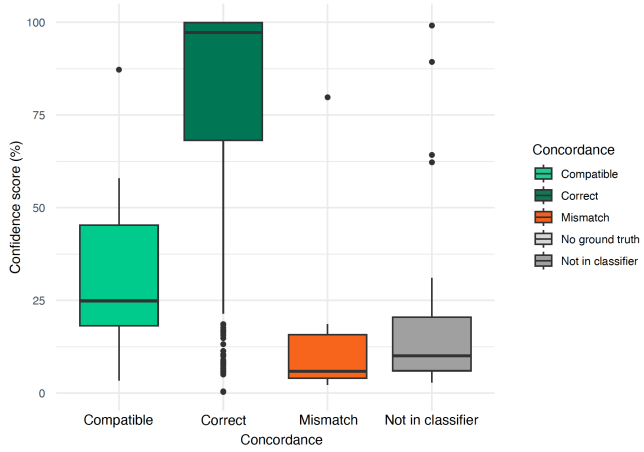


Figure 3-17 Distribution of confidence scores for Rapid-CNS^2 methylation classification

Another three samples were assigned to a reactive tumour microenvironment category, correctly identifying the context of the tumour but not specifying the precise tumour type. The remaining 8 out of 15 non-matching samples were considered clear mismatches. Notably as shown in **Figure 3-17**, samples that were correctly classified tended to have higher confidence scores than those that were mismatched or outside the reference set, indicating a generally conservative approach. An important challenge for classification models is the handling of entities not present in the current classifier reference set. As demonstrated in **Figure 3-17**, all cases belonging to classes not in the current classifier (eg. metastases) were predicted with low calibrated scores. In contrast, samples with concordant methylation classes had high calibrated scores.

3.3.10 Copy number variation calling and validation

Copy number calls made by the Rapid-CNS^2 pipeline were compared to their corresponding EPIC array and NGS panel sequencing analysis results. Visual inspection was carried out for comparison as it is the standard practice for evaluating such results in neuropathology diagnostics. Complete concordance in copy number profiles with EPIC array was observed for all samples. CNV profiles generated by Rapid-CNS^2 displayed improved resolution over those from NGS panel sequencing for the same targets (Panel B) as shown in **Figure 3-18**. This can be attributed to the rejected reads that are uniformly distributed across the genome.

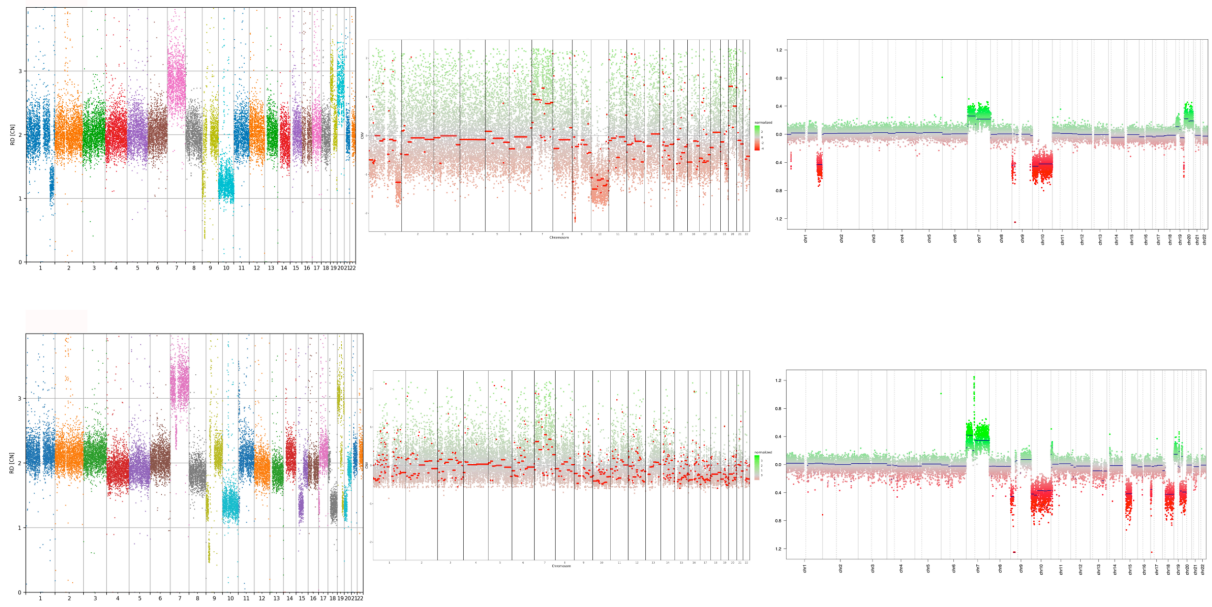


Figure 3-18 Copy number profiles from Rapid-CNS^2 (left), NGS panel sequencing (middle) and EPIC array (right)

Reducing the panel size did not reduce the resolution of copy number calls. Samples targeting Panel A and B had similar accuracy of copy number variation detection. Since the copy number profile is primarily results from the rejected short reads, it follows that the copy number profiles should remain similar regardless of panel size.

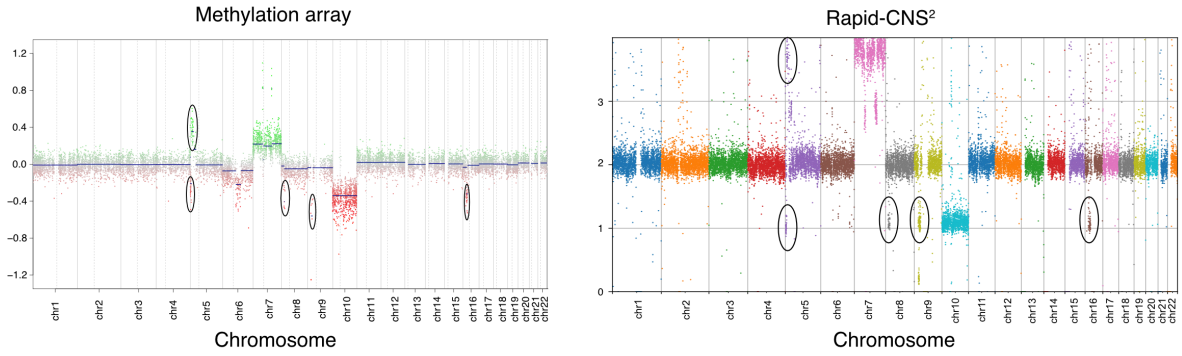


Figure 3-19 Concordance of focal alterations in CNV profiles. (Reprinted from Patel et. al 2024[117])

On account of their pathognomonic nature, arm-level copy number alterations like 1p/19q codeletion and 7 gain/10 loss were particularly investigated. In all sequenced glioma samples, these alterations were detected with 100% accuracy.

Focal alterations like *EGFR* amplification were reliably detected in all cases. Visual inspection of the CNV profiles from EPIC array analysis is the standard practice for diagnostic decisions

and was also used as the ground truth for focal deletions. **Figure 3-19** shows an example of a glioblastoma sample with multiple focal alterations that were accurately identified by Rapid-CNS^2. Copy number calculations by Rapid-CNS^2 also report genes falling within variant segments. This was additionally considered while assigning gene-level copy numbers. Rapid-CNS^2 considers multiple bin sizes (1kb, 10kb and 100kb) for copy number calling. While not reported in this work, a combination of small and large bin sizes could further contribute to the reliable calculations of gene and arm-level copy numbers.

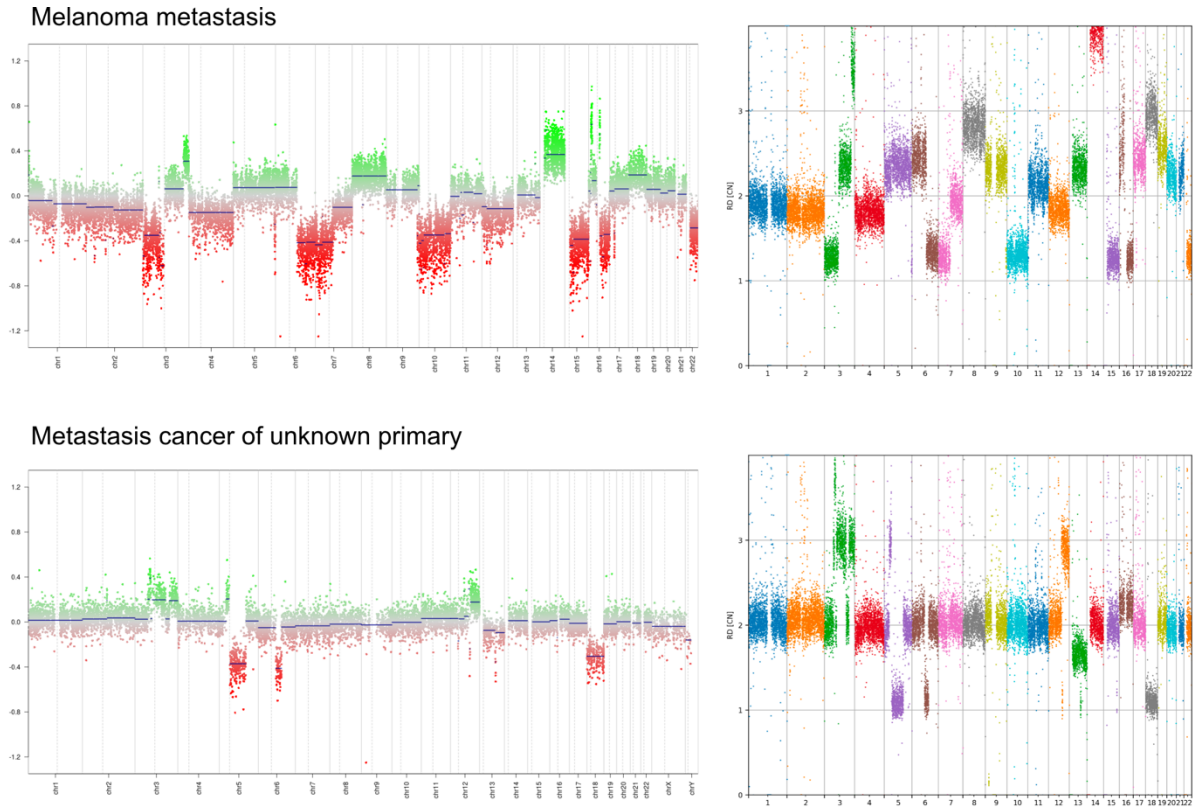


Figure 3-20 Comparison of CNV profiles for brain metastases samples

Metastatic tumours are known to have a host of complicated genome-wide copy number alterations. I compared CNV profiles for such brain metastases samples and found them to be consistent with their respective CNV profiles from EPIC array analyses. **Figure 3-20** shows the copy number profile for metastases samples by methylation array analyses on the left and that generated by Rapid-CNS^2 on the right. Focal alterations in the samples were concordant and could be detected with high resolution.

Copy number variants are widely known to be of prognostic and even pathognomonic value in CNS tumours[13,143,165,166]. The list of relevant copy number alterations is constantly increasing. For example, the recent c-IMPACT-NOW update 8 additionally proposed to assign a higher WHO grade to meningiomas with WHO grade 1 histology but chromosome 1p arm deletion in combination 22q deletion[167]. It has recently been suggested that size-dependent CNVs significantly influence risk stratification in cancer genomes[168]. While methylation arrays can indicate arm and gene level alterations, they are dependent on individual probes over the genome which might affect the detection of such sizes. On the other hand, sequencing reads cover entire regions enabling finer resolution and more precise size detection of CNVs. With the added benefit of SNP/SNV detection, sequencing data provides the additional capability to assess loss of heterozygosity (LOH) and allele-specific sizes.

Copy number profiling emerged as one of the strongest features of the Rapid-CNS^2 approach, with results demonstrating complete concordance with methylation array analyses at both chromosomal and focal levels. This success can largely be attributed to adaptive sampling, which rejects 90% of off-target reads within 500 base pairs, allowing the remaining reads to effectively span diverse genomic regions. This precise targeting was particularly valuable in cases where methylation classification failed, as copy number profiles played a key role in providing molecular diagnoses. However, it is important to note that in this study, the nature of nanopore-based copy number variant (CNV) results differed significantly from those generated by the EPIC array, making it impossible to quantify concordance by directly comparing scores. This highlights the need for a more systematic approach to quantify CNV concordance between different platforms, particularly in assessing zygosity, since zygosity of *CDKN2A/B* for example is pathognomonic for CNS tumours. Additionally, cut-offs beyond simple visual inspection should be defined to determine the zygosity of CNVs. Clear thresholds for distinguishing between heterozygous and homozygous deletions would further enhance the diagnostic precision of CNV profiling in nanopore-based workflows. Developing such quantitative metrics would improve the reliability of CNV analysis, making it a more valuable component of the diagnostic toolkit.

3.3.11 Structural variant calling

I focused the validation of fusion calls on clinically relevant fusions present in the “ground truth” samples. svim and Sniffles (or Sniffles2) detected a range of long structural variants including

deletions, inversions and duplications[138,139,169]. Since the matched NGS data was based on panel sequencing, it covered only targeted regions of the genome. In contrast, targeting through adaptive sampling retains the entirety of the reads mapping to the regions of interest. More importantly, long reads reliably represent the entire landscape of a structural variant whereas short reads require considerable amount of inferencing to unravel a long structural variant. Due to the limitations in reliable SV calling from short read panel sequencing data, no ground truth was available to validate SVs called by Rapid-CNS[2].

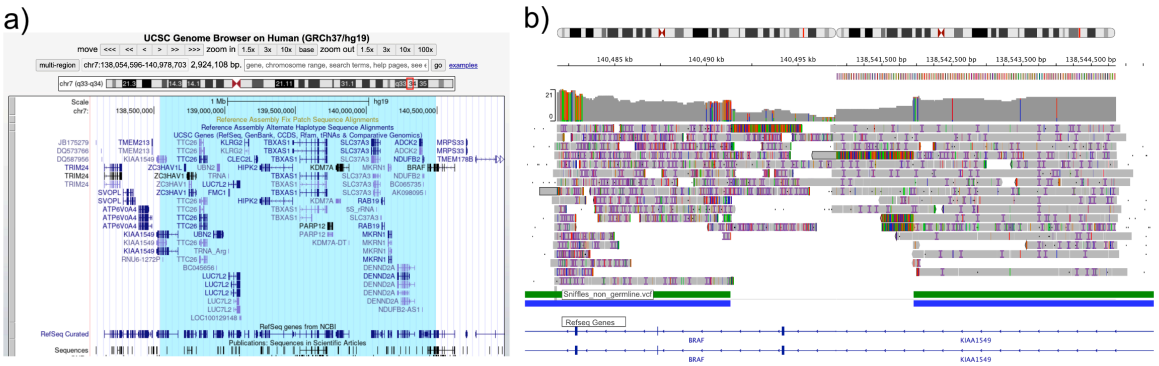


Figure 3-21 *BRAF:KIAA1549* fusion detection (Reprinted from Patel et. al 2024[117])

I specifically queried fusions from the SV calling results depending on the suspected diagnosis or predicted methylation class. For example, I looked for *BRAF:KIAA1549* fusions in pilocytic astrocytoma cases and CIC rearrangements in a sample predicted as EFT_CIC (*CIC* altered sarcoma) by methylation. I found *BRAF:KIAA1549* fusions in all 7 cases with the fusion identified in corresponding NGS-RNA seq data. Interestingly, the fusion was called as a 19Mb duplication on chr7.q34 with breakpoints located in the introns of *BRAF* and *KIAA1549* genes. The duplicated region is highlighted in blue in **Figure 3-21** a). This is in line with the tandem duplication based mechanism of the fusion proposed by Jones et. al[170]. IGV screenshot of the breakpoints is shown in **Figure 3-21** b). Bars at the bottom indicate duplication region annotated by Sniffles[138].

Sample 173 was predicted as a *CIC* altered sarcoma by methylation. This diagnosis warrants presence of a *CIC* (capicua transcriptional repressor) fusion usually with *LEUTX* (leucine twenty homeobox), *NUTM1* (NUT midline carcinoma family member 1), or *DUX4* (double homeobox 4) family genes[171-174]. No fusion was detected by NGS DNA or RNA sequencing. I found a single 13kb read, with 3kb mapping to exon 20 of the *CIC* gene (chr19) and 10 kb

mapping to the region of *DUX4* retrogenes (chr4) specifically to *DUX4L5*. The read is highlighted in orange in **Figure 3-22**.

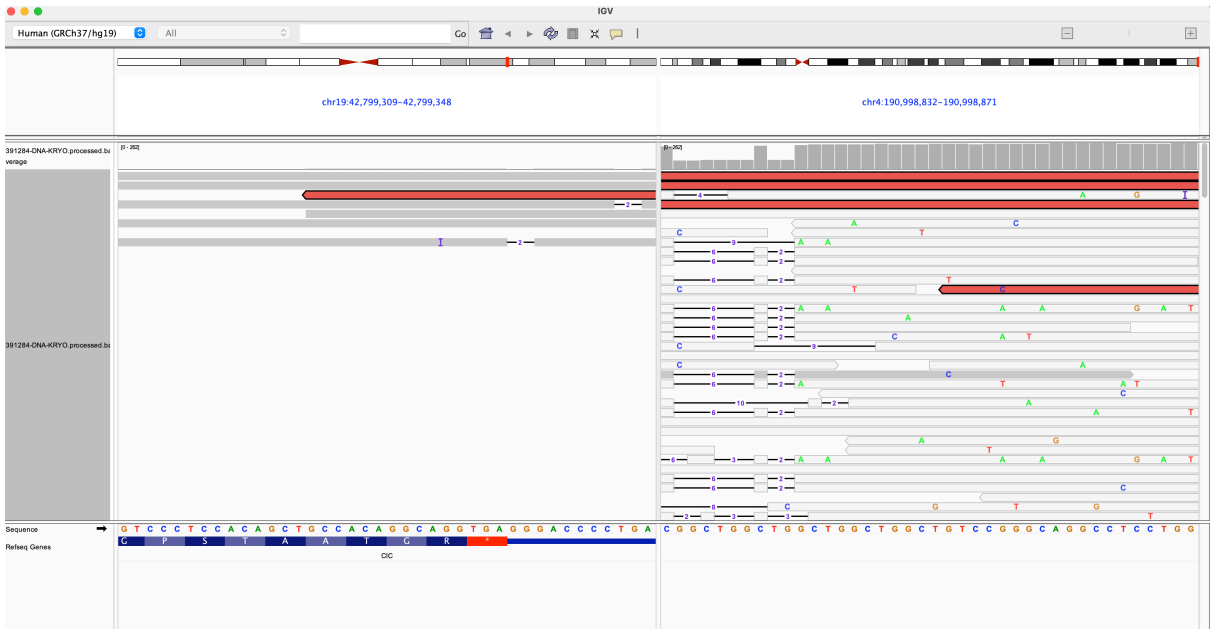


Figure 3-22 IGV screenshot for *CIC* exon 20 (chr19) and *DUX4* retrogene (chr4) regions

While fusions with *DUX4* retrogenes on chr4.q35 have been reported widely, these have been detected by short read sequencing by aligning to the hg19 genome. In the hg19 genome, the *DUX4* gene is present on an unplaced scaffold and the retrogenes are on chr4, which was resolved and then placed next to the retrogenes on chr4 in the hg38 genome[121,133]. This implies that effective alignment of reads to this gene in hg19 would be challenging. This demonstrates the need to incorporate newer reference assemblies for better resolution and fusion detection. I did a BLAST (basic local alignment search tool) comparison of the *DUX4L5* gene to the *DUX4* gene and found that the retrogene had a 99.7% match to exon 1 of the *DUX4* gene[175,176]. The usual location of the *DUX4* gene fusion breakpoint is in exon 1. Since short reads only cover up to 300 bp which is shorter than an exon, the exon 1 breakpoint detected for *DUX4* by NGS could in fact be from one of the retrogenes. I speculate that long read sequencing would better resolve the actual rearrangement since they cover bases to the order of kilo- or even megabases.

Long read sequencing offers a major advantage over NGS panel sequencing in being able to resolve long complex variants in a single read. Due to unavailability of ground truth data, SV calling for this work was restricted to purely exploratory analysis.

A 1.3 Mb deletion in *EGFR* (epidermal growth factor receptor) spanning the exons 2 through 7 was called by Sniffles2 in one glioblastoma sample as shown in **Figure 3-23**. Deletion of exons 2-7 is characteristic of the *EGFR* vIII variant. The deletion was found in 18 out of 100 reads covering the region, thus making it a subclonal variant. All 18 reads spanned both the breakends and extended over 300 bp on either side. The read highlighted in orange indicates the same read. To confirm the variant, the sample was sequenced again using an R10 flowcell. The variant was not detected in NGS panel sequencing.

Figure 3-23 *EGFR* vIII detection (Reprinted from Patel et. al 2024[117])

Interestingly, the *EGFR* gene was amplified in this sample without a chr7 gain. Since the variant was found in ~20% of the reads covering the gene, it follows that the variant was acquired after the amplification. This is consistent with clonal evolution analyses of *EGFR* in glioblastomas that deem *EGFR* amplification an early clonal event that is followed by acquisition of *EGFR* vIII mutations as a result of intratumoural heterogeneity[177-179].

3.3.12 Integrated diagnoses

Integrated diagnoses were issued considering histology, clinical data and molecular data reported by Rapid-CNS[2]. **Figure 3-24** shows the concordance over the multiple layers of evaluation for the entire cohort. While methylation classification is a major criterion, Rapid-CNS[2] also possessed the added advantage of reporting SNVs, CNVs and gene fusions. These

layers contributed significantly to cases where methylation classification was incorrect or unclassifiable, and to cement diagnoses that cannot be made based on methylation classes and histology alone.

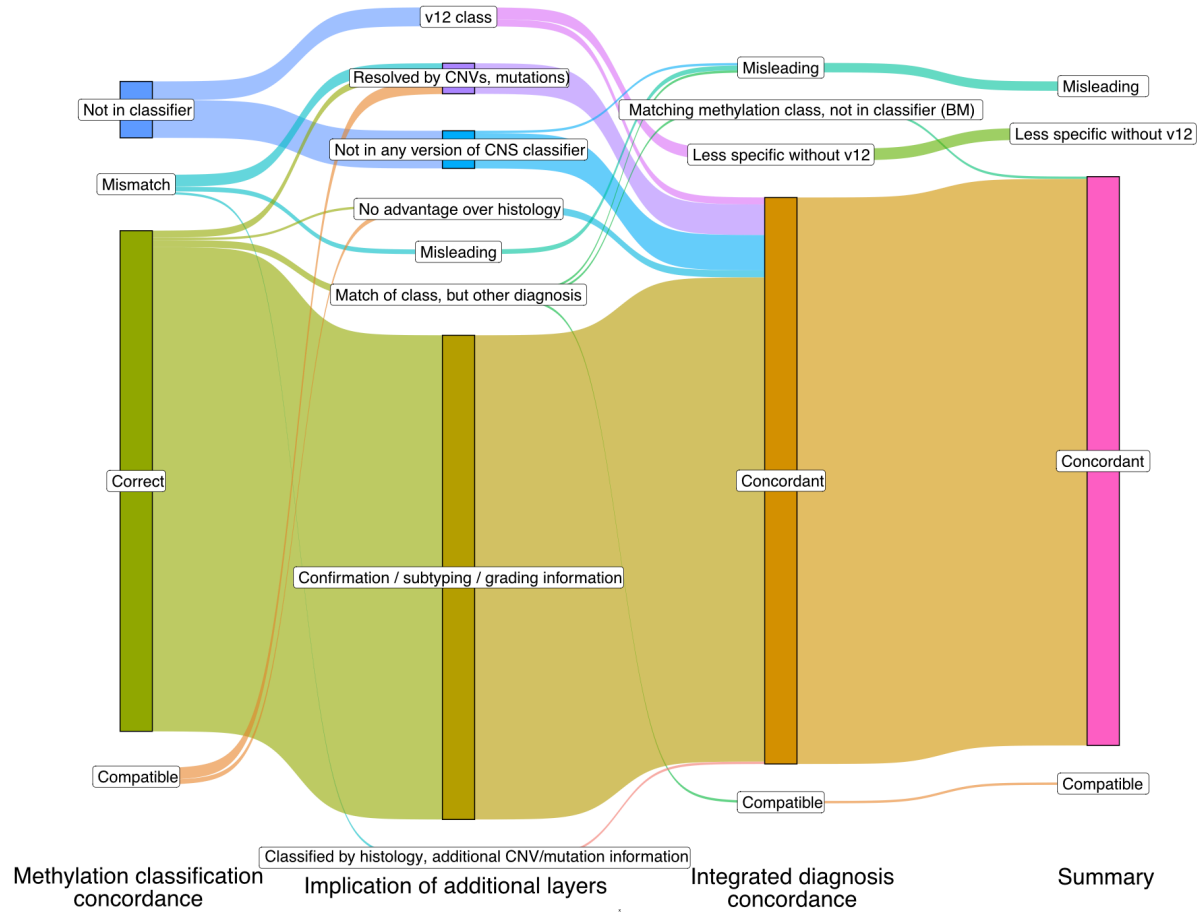


Figure 3-24 Concordance over layers of evaluation of Rapid-CNS^2 results for the cohort

I found integrated diagnoses to be in concordant in 242 out of 251 cases. One case was issued a diagnosis compatible with the conventional integrated diagnosis (Sample 124). The EPIC array methylation family for the sample matched the Rapid-CNS^2 prediction albeit both of them had very low scores. Conventional analysis discovered an *IDH1* R132H mutation and mutations indicating MMR-deficiency, leading to an integrated diagnosis of 'Primary mismatch repair deficient *IDH*-mutant astrocytoma'. On the other hand, Rapid-CNS^2 reported a hypermutant phenotype with mutations in *MSH6* but no *IDH1* mutation and a flat copy number profile. This led to an integrated diagnosis of 'Mismatch repair deficient glioma', compatible with the conventional diagnosis. Five samples suffered from lack of specific class in the Rapid-CNS^2 classification scheme (MNP v11). EPIC array had an advantage in these case since it

employs the v12 classifier with 184 classes. While four of these tumours could be assigned a coarse diagnosis, one of them remained unresolved. Only four cases had potentially misleading diagnoses (1.6%). Importantly, none of them presented a consistent picture of a confidently but incorrectly called diagnosis across all layers, but evidently called for additional analysis in keeping with the integrated diagnosis concept of the WHO classification.

3.3.13 Overall concordance

To thoroughly evaluate the performance of Rapid-CNS^2 on a highly diverse, nearly 'come as you are' basis, we prospectively included cases that, in hindsight, might not have been suitable for methylation-based analysis (e.g., non-primary brain tumours not represented in the CNS methylation classifier). However, these cases were not excluded from the final analyses to provide a complete assessment of the capabilities of Rapid-CNS^2. Even in cases where methylation was uninformative, other data layers such as CNVs, mutations, and fusions still contributed significantly to the diagnostic process. The added value beyond methylation classification in rendering a precise diagnosis encompassed pathognomonic CNVs (e.g. 7/10 in glioblastoma, 1p/19q in oligodendroglioma) and/or pathognomonic mutations (e.g. *IDH1*, *TERT*) and/or gene fusions (e.g. *BRAF*:*KIAA1549*) to distinguish between differential diagnosis in cases not resolved by methylation alone. Remarkably, all small biopsy, recurrence and infiltration zone samples could be issued concordant integrated diagnosis. **Figure 3-25** and **Figure 3-26** show the overall concordance for archival and diagnostic samples including intraoperative samples respectively.

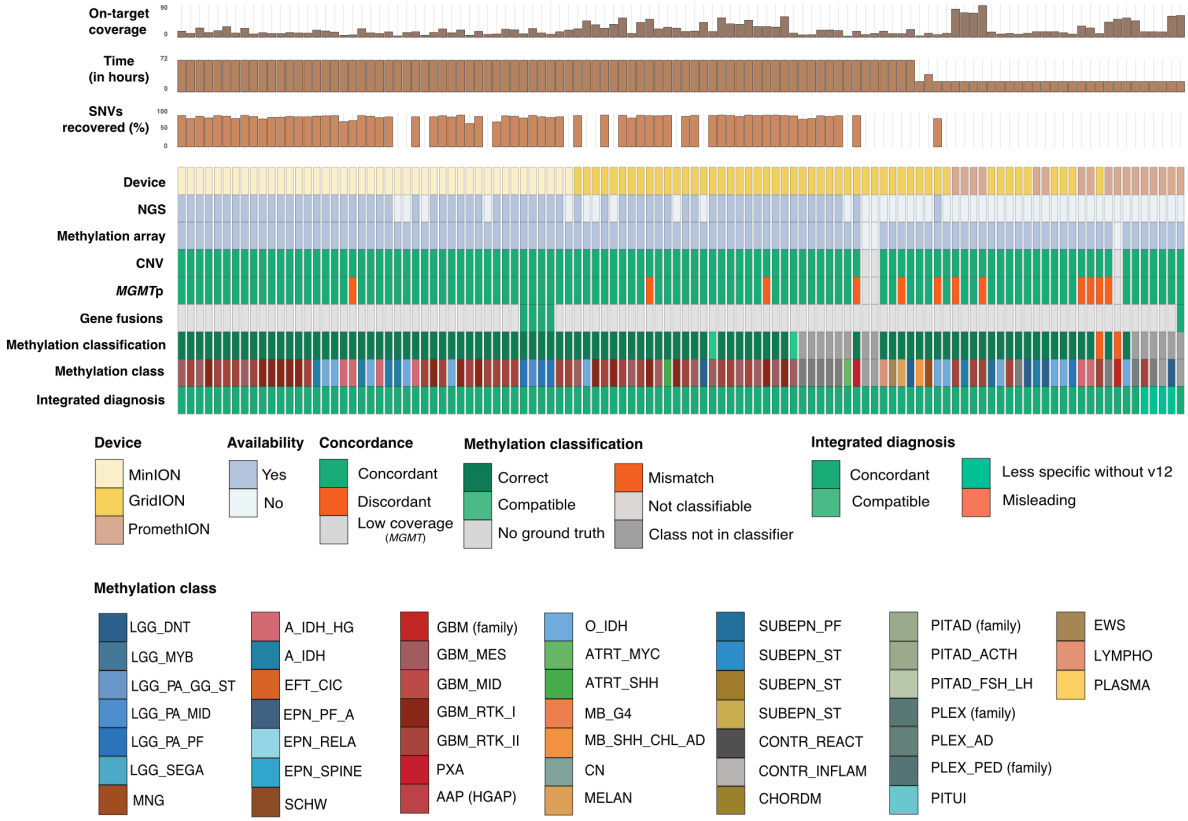


Figure 3-25 Overview of archival samples

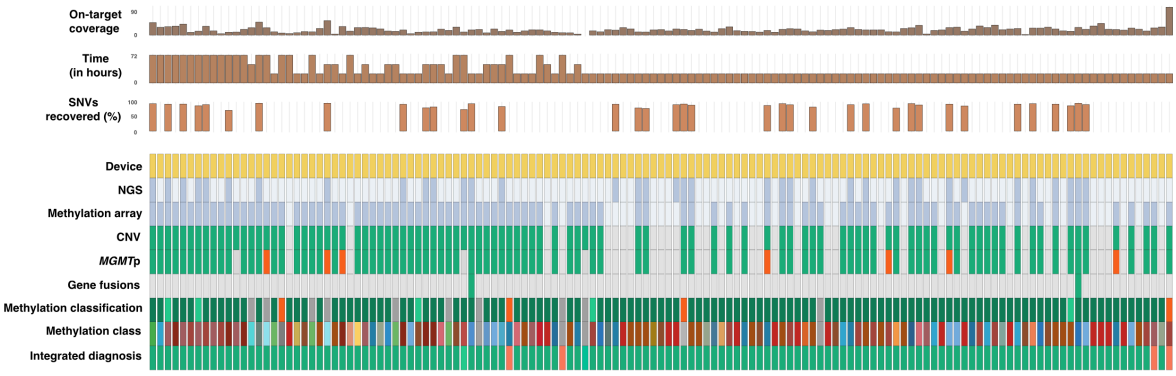


Figure 3-26 Overview of diagnostic samples

3.3.14 Intraoperative sequencing

This extract was adapted from Patel et. al 2024[117]:

"To evaluate the shortest time needed for methylation classification and copy number calling, I re-analysed data from 36 representative samples taken from intra-operative frozen sections in the Heidelberg Rapid-CNS^2 dataset, which had been previously run on R9 flowcells. These samples had originally been analysed from archival tissue, so I recreated a real-world sequencing scenario by sub-setting the data to include only reads generated at various time points during the actual run. I used the sequencing summary file generated by guppy to identify the reads. **Figure 3-27** displays how methylation class assignment progressed over time. After merely 15 minutes of sequencing, 29 out of 35 (83%) samples with sufficient reads were assigned the correct methylation family. After an hour, 35 out of 36 samples were accurately classified. Importantly, concordant classifications had significantly higher calibrated scores than discordant ones at all time points as indicated by the boxplots in the top plot of **Figure 3-27**.

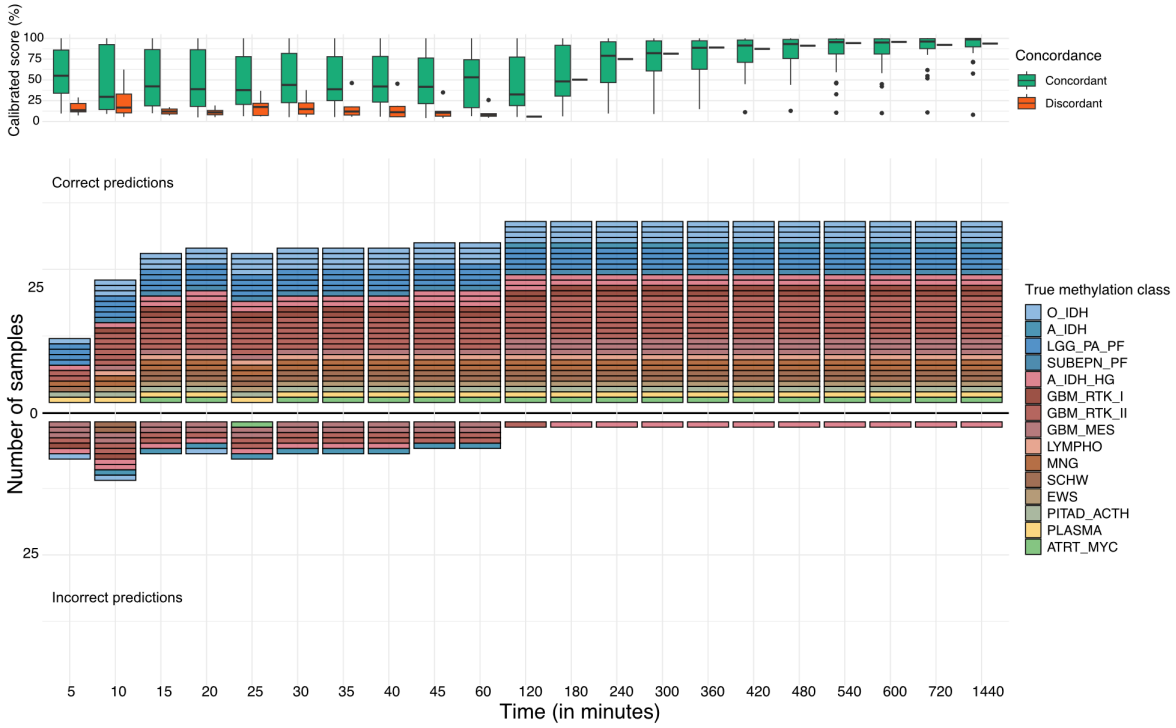


Figure 3-27 Simulated intraoperative methylation class reporting. (Reprinted from Patel et. al 2024[117])

Adaptive sampling's rejected short reads led to uniform genome coverage, allowing for high-resolution copy number profiles within short timeframes. Arm-level alterations, such as the loss of 1p, 7p, and 22q, were clearly detected after only 10 minutes of sequencing. Notably, one sample identified as an *IDH*-wildtype glioblastoma based on copy number alterations, including diagnostic chromosome 7/10 changes, had initially been misclassified as a 'high-grade *IDH*-mutant astrocytoma' via methylation profiling until after 24h of sequencing."

The success of the simulated experiments prompted the establishment of true intraoperative experiments. Rapid protocols for intraoperative methylation classification have been reported in multiple studies before[163,164]. These protocols are based on shallow whole genome sequencing, which while swift, are restricted to detecting methylation classification and lower resolution copy number alterations. Matthew Loose's lab devised a protocol that combined the ultra-long kit from ONT with the rapid sequencing kit to reduce the DNA extraction and library preparation time to less than 60 minutes from 7 hours. This protocol was improved by Jochen Meyer to further reduce the time to 50 minutes. We performed sequencing on a P2 Solo using adaptive sampling by ReadFish in Nottingham and Heidelberg[122,123]. This led to a workflow that started intraoperatively with methylation classification and broad CNV reporting within 30

minutes followed by a full spectrum of alterations reported after 24h of sequencing as demonstrated in **Figure 3-28**.

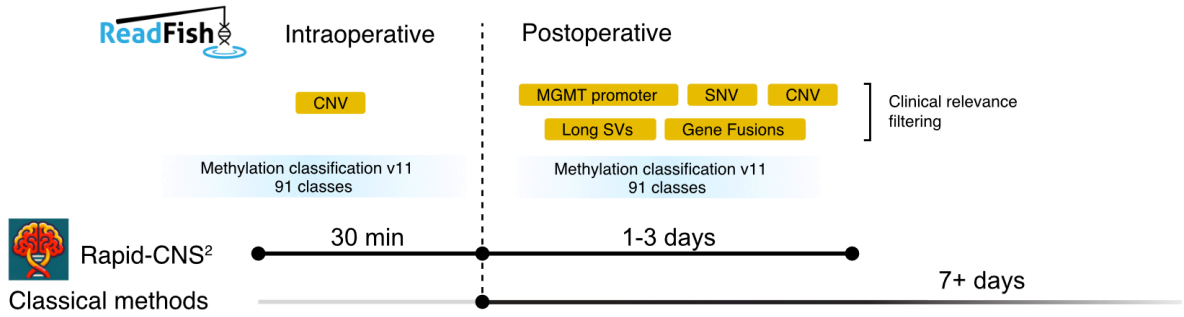


Figure 3-28 Schematic of intraoperative Rapid-CNS^2 pipeline. (Reprinted from Patel et. al 2024[117])

Using this modified rapid library preparation protocol, we conducted real-time intraoperative sequencing on a combined 18 samples in both centres. **Figure 3-29** shows the predictions over an hour of sequencing for 16 of the classifiable samples, with each centre running its analysis independently. Consistent with my retrospective simulations, integrating real-time methylation and CNV data improved the precision of interpretation for 13 out of 18 intraoperative samples using Rapid-CNS^2, with results available within 30 minutes. One of the unresolved samples was a novel entity unclassifiable with the methylation classes available and the other was not classifiable by the methylation array. Importantly, these two samples were consistently predicted with very low confidence (<25%), supporting a conservative approach taken by the classifier. The samples predicted confidently were relayed to the surgeon in the operating room by phone.

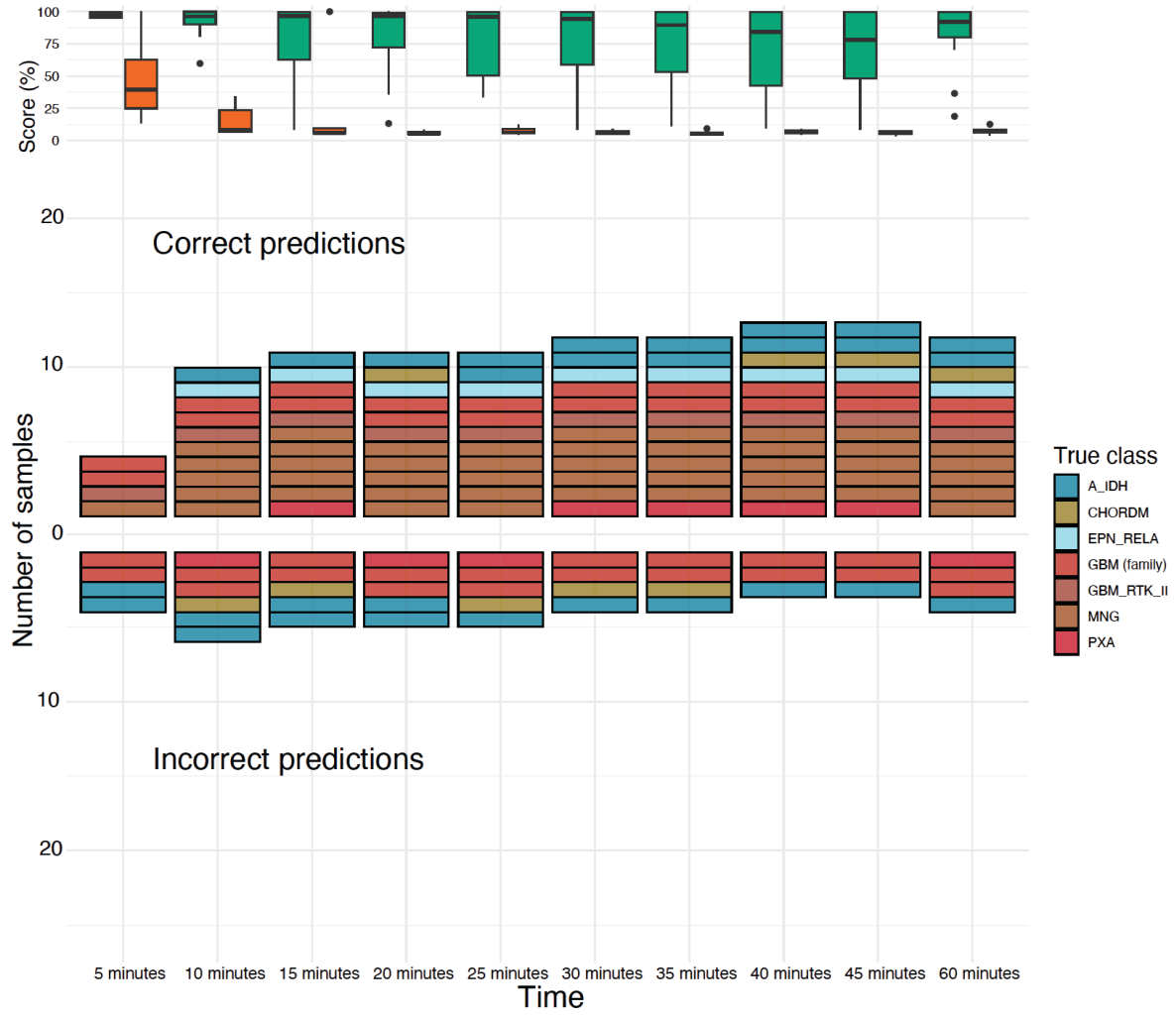


Figure 3-29 Methylation classification results over time for real intraoperative sequenced samples. (Reprinted from Patel et. al 2024[117])

Since I only had access to the aligned bam files from the Heidelberg data, I also analysed the CNV profiles from these samples. On the other hand, I only analysed the methylation data for the samples processed in Nottingham.

Table 3-2 describes the results for the samples sequenced intraoperatively in Heidelberg. Methylation class predictions for three out of five samples were completely concordant with the final conventional integrated diagnosis after only 30 minutes of sequencing. Notably, their copy number profiles were also accurately represented in the intraoperative timeframe.

Table 3-2 Heidelberg intraoperative sequencing results

ID	*Rapid-CNS² classification*	*Rapid-CNS² CNV*	*Final conventional integrated diagnosis*	*Comment*
248	No clear classification, consistent PXA classification from 35 minutes onwards	chr7 gain / chr10 loss estimable from 5 minutes onwards	Gliosarcoma, IDH-wildtype, WHO grade 4	Recurrence post therapy, no clear score with methylation array
249	No clear classification, consistently low scores (<10%)	chr7 gain / chr10 loss	Glioblastoma, IDH-wildtype, WHO grade 4	Resolved by CNV. Poor tissue selection, library preparation.
250	A_IDH from 20 minutes onwards	Segmental gains in chr7, chr8; segmental losses in chr3, chr4, chr19	Astrocytoma, IDH-mutant, WHO grade 2	Concordant with final diagnosis
251	A_IDH from 30 minutes onwards	Flat copy number profile	Astrocytoma, IDH-mutant, WHO grade 2	Concordant with final diagnosis
252	GBM_RTK_II from 5 minutes onwards	chr7 gain / chr10 loss estimable from 5 minutes onwards	Glioblastoma, IDH-wildtype, WHO grade 4	Concordant with final diagnosis

Of the two unclassifiable samples, ID 248 was the recurrence of a glioblastoma. As evident in **Figure 3-30**, this sample gave inconsistent classifications with very low scores until one hour of sequencing. Even after 24h of sequencing, the methylation class prediction was pleomorphic xanthoastrocytoma (PXA). There was no methylation array for this resection. Methylation array analyses of earlier resections gave a score of 0.32 for the GBM_MES class, deeming the sample unclassifiable. But, there was an evident 7/10 signature from 10 minutes of sequencing onwards, leaning the intraoperative diagnosis towards recurrent glioblastoma. The histology description from the FFPE section described an abundance of *CD68* positive cells (most likely macrophages). Post-therapy recurrent glioblastoma to gliosarcoma transition is known to have increased immune cells, reactive astrocytes and macrophages that can impair methylation classification[180,181]. However, the intraoperative CNV profile was consistent with that generated by the methylation array analyses of previous resections and stayed similar even after 24h of sequencing. It was nearly identical with segmental 3p loss, 7/10 signature, 9q gain, segmental 13q loss, *NF1* deletion, and gain of chromosome 20. While there exists an overlap between PXA and GBM (for e.g. 7/10 signature, *CDKN2A/B* deletion),

the diffuse nature of the tumour, absence of *BRAF* V600E mutation and early recurrence with evident sarcomatous phase on histology, led to the final diagnosis of gliosarcoma in concordance with conventional analysis.

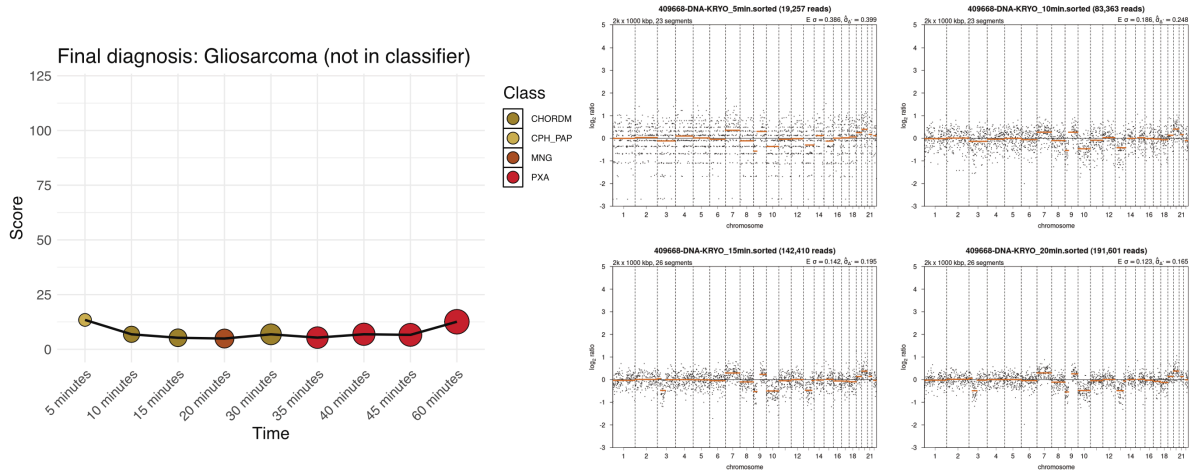


Figure 3-30 Intraoperative classification and copy number profiles for gliosarcoma sample

The second sample (ID 249) had poor tissue selection leading to suboptimal library preparation. This is evident from the number of reads generated over time, the sample only accumulated 77,000 reads after 20 minutes compared to 191,000 reads for the previous sample (ID 248) at the same timepoint. This was reflected in the low scoring, inconsistent methylation classification over time for the sample as depicted in **Figure 3-31**. However, the copy number profiles showed a clear 7/10 signature from 5 minutes onwards, leading to a glioblastoma diagnosis in line with the conventional diagnosis established later.

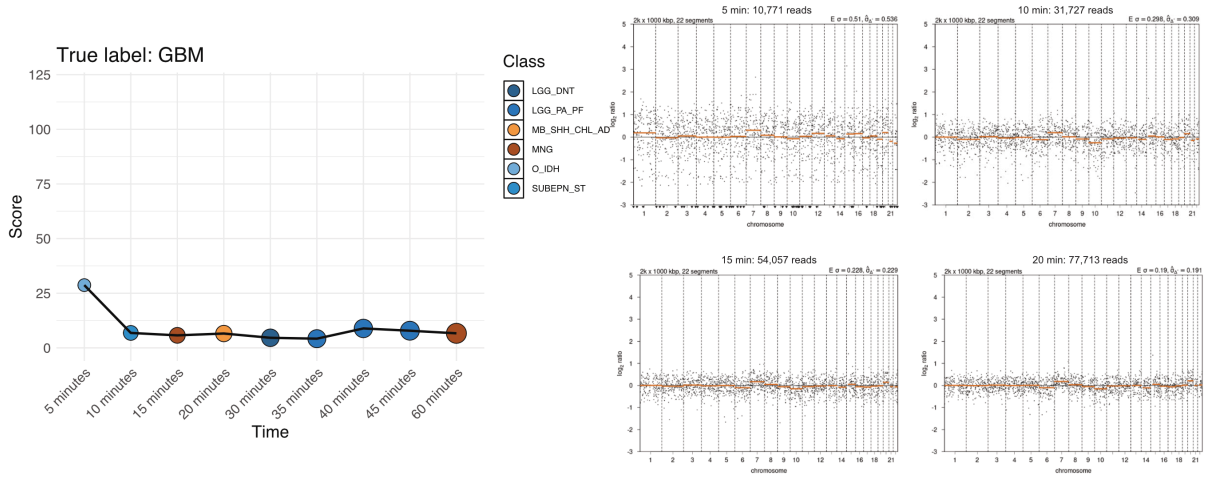


Figure 3-31 Intraoperative sequencing resolved by CNVs for glioblastoma (GBM) (Reprinted from Patel et. al 2024[117])

As presented in **Table 3-3**, eight out of eleven classifiable samples from Nottingham were correctly classified intraoperatively. Two samples (Nott_6 and Nott_13) were unclassifiable as mentioned above. I was able to accurately identify three diffuse glioma samples as either *IDH*-mutant astrocytoma or *IDH*-wildtype glioblastoma, diagnoses that are difficult to distinguish based on frozen section alone. Similarly, one ependymoma sample was classified as part of the high-risk supratentorial *ZFTA*-fusion positive group, differentiating it from the lower-risk *YAP1*-fusion diagnosis, which is histologically similar. I could also accurately predict five meningioma cases with high scores from 10 minutes onwards. Only three cases offered no substantial improvement over standard morphological inspection of frozen sections, and one of these remained unresolved even after conventional testing, labelled simply as a "glial neoplasm." Importantly, these three cases were constantly predicted with low scores, consistent with similar observations in the Heidelberg data.

Table 3-3 Nottingham intraoperative sequencing results

ID	*Rapid-CNS² classification*	*Final conventional integrated diagnosis*	*Comment*
Nott_1	EPN_RELA from 10 minutes onwards	Supratentorial ependymoma, ZFTA fusion-positive	Concordant with final diagnosis
Nott_2	GBM_RTK_II from 5 minutes onwards	Glioblastoma, IDH-wildtype, WHO grade 4	Concordant with final diagnosis
Nott_3	MNG from 5 minutes onwards	Meningioma	Concordant with final diagnosis

Nott_4	MNG from 10 minutes onwards	Meningioma	Concordant with final diagnosis
Nott_5	Consistently misclassified with low scores	Astrocytoma, IDH-mutant, WHO grade 2	Low tumour DNA fraction
Nott_6	CONTR_INFLAM from 10 minutes onwards	Glial Neoplasm, NOS	Not classifiable with array
Nott_7	GBM_RTK_II from 5 minutes onwards	Glioblastoma, IDH-wildtype, WHO grade 4	Concordant with final diagnosis
Nott_8	MNG from 5 minutes onwards	Meningioma	Concordant with final diagnosis
Nott_9	MNG from 5 minutes onwards	Meningioma	Concordant with final diagnosis
Nott_10	MNG from 10 minutes onwards	Meningioma	Concordant with final diagnosis
Nott_11	CHORDM from 40 minutes albeit with low scores	Chordoma	Poor tissue selection
Nott_12	PXA from 30 minutes with low scores	Pleomorphic xanthoastrocytoma, CNS WHO 3	Necrosis & microvascular proliferation
Nott_13	Consistently misclassified with low scores	Diffuse Paediatric-type high grade glioma, RTK1, subclass A	Novel entity, not classifiable

Overall, intraoperative Rapid-CNS^2 provided clinically relevant information on tumour subtype and risk profile in 72.2% of cases. The intraoperative protocol yielded results on tumour classification and CNVs within 90 minutes of sample receipt, with sequencing and data interpretation taking just 30 minutes. I provided a comprehensive report, including SNVs, indels, gene fusions, and detailed methylation classification, the following day.

3.3.15 Routine application

Once the pipeline was fully implemented at the Department of Neuropathology, UKHD for routine prospective use, 51 out of 62 CNS tumour samples (82.3%) received from February to May 2024 were processed through the Rapid-CNS^2 workflow without tissue quantity or quality limitations. I investigated the potential advantages of intraoperative sequencing in these samples by simulating real-time analysis and comparing results after 30 minutes of sequencing. As is clear from **Figure 3-32**, most cases could be issued granular molecular diagnoses within the intraoperative window concordant to the final integrated diagnoses and

a major improvement from the broad histology-based diagnoses usually issued in the same timeframe.

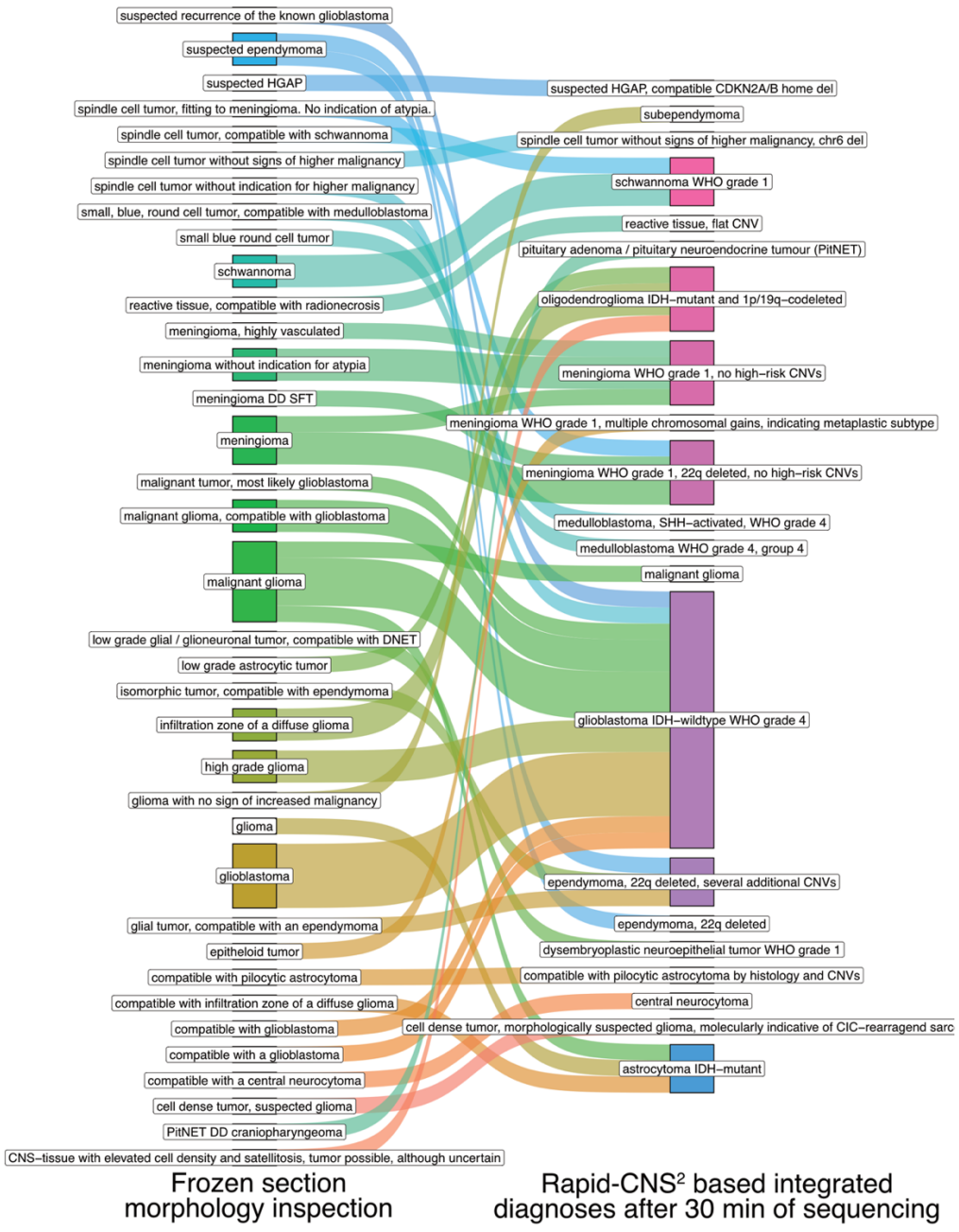


Figure 3-32 Improvement over frozen section histology diagnoses after 30 minutes of sequencing (Reprinted from Patel et. al 2024[117])

In this dataset, Rapid-CNS^2 demonstrated a high level of concordance with conventional diagnostic methods, offering results potentially within an intraoperative timeframe compared

to the 15-20 days typically required for traditional diagnostics. As shown in A.1.3, out of the 51 cases analysed, 15 out of 16 cases were consistently diagnosed as Glioblastoma *IDH*-wildtype WHO grade 4 by both Rapid-CNS^2 and conventional methods. Methylation classification was correct in 14 cases, while one cases with a wrong classification could be resolved by histology and a 7/10 signature in the copy number profile. One case (ID 200) was issued a reactive tissue diagnosis as per its histology, inconsistent methylation classification and a flat CNV profile. One gliosarcoma (ID 184) was diagnosed as a glioblastoma by 7/10 signature in the CNV profile of Rapid-CNS^2, which is a compatible diagnosis. This would have been crucial to know since the histology-based frozen section diagnosis of this sample was that of a spindle cell tumour without indication for higher malignancy. This reliable early confirmation of these aggressive tumours could allow for prompt surgical decisions aimed at maximising resection margins. 9 meningioma cases were identified as WHO grade 1 tumours with no high-risk CNVs through nanopore sequencing. The methylation class was concordant in 8/9 with the conventional diagnosis, with one discrepant tumour being identified by histology and 22q deletion. Early identification of low-risk meningiomas would enable surgeons to take a more conservative approach by avoiding overly aggressive procedures for these slow-growing tumours. Three astrocytomas were correctly identified of which two were also accurately classified as high grade. For ID 219, methylation classification indicated suspected diffuse midline glioma (DMG_K27) with low score, but histology and *CDKN2A/B* homozygous deletion were suggestive of a high-grade astrocytoma with piloid features (HGAP), compatible with the final diagnosis. 4 oligodendroglioma cases were also accurately diagnosed intraoperatively, identified by either methylation class and/or the 1p/19q codeletion. One case (ID 209) could not be classified and was predicted as a malignant glioma which was finally described as a NEC glial/glioneuronal tumour with *VHL* mutation by conventional analysis. Additionally, 1 case of desmoid-type fibromatosis (ID 175) was could also not be classified by methylation and was diagnosed as a spindle cell tumour with a chromosome 6 deletion, which was later confirmed by conventional diagnostics. This tumour was classified as desmoid-type fibromatosis by the sarcoma classifier in conventional analysis, a tool not available for nanopore methylation data. Though benign, this tumour is locally aggressive, and early identification via nanopore sequencing could have enabled accurate classification of this rare tumour type. This calls for more classifiers to be trained for nanopore data. Two medulloblastomas (IDs 185, 203) were diagnosed as WHO grade 4 tumours, with one being *SHH*-activated (ID 185) and the other from group 4 (ID 203). Given the rapid progression and poor prognosis of medulloblastomas in paediatric cases, intraoperative molecular diagnosis

offers substantial advantages for resection strategy as well as initiating timely treatment plans, including postoperative radiation and chemotherapy. Another complex case was ID 173, where frozen section analysis initially suggested a glioma. Nanopore sequencing, however, classified it as a *CIC*-rearranged sarcoma ('EFT_CIC'), a rare but aggressive tumour. Conventional methods later confirmed this diagnosis, although RNA sequencing failed to detect a *CIC*-fusion. This case emphasises the potential of the workflow to reveal molecular signatures even when traditional techniques are inconclusive.

3.3.16 Improvements over conventional methods

While intraoperative sequencing can influence the extent of resection for CNS tumours, Rapid-CNS^2 run postoperatively also majorly improves over conventional methods. The turnaround time is reduced to 2 days compared to over 20 days for conventional molecular analysis. In addition, long reads enable detection of single nucleotide and structural variants in addition to methylation. This could potentially replace methylation array, NGS DNA panel sequencing and NGS RNA sequencing, not to mention whole genome sequencing that might be needed if no relevant variants are detected by any of the previous methods. For example, sample 173 was issued a frozen section diagnosis of ‘glioma’ based on the smear and frozen section (**Figure 3-33**).

This sample was sent for Rapid-CNS^2 analysis with results available within 5 days without any prioritisation. The sample had a methylation classification of Ewing-family tumour, *CIC* altered (EFT_CIC) with a high confidence score. As discussed in the previous section, this diagnosis was already available within 30 minutes of sequencing. The copy number profile indicated a chr6 gain. As suggested by the name, this diagnosis warrants identification of *CIC* rearrangement in the sample. Rapid-CNS^2 detected a *CIC:DUX4L5* fusion which was not detected by NGS DNA or RNA sequencing which were eventually available over 25 days later (**Figure 3-22**).

This case demonstrates the multifaceted advantages of Rapid-CNS^2- it accurately reported the methylation class, chr6 deletion and the *CIC* gene fusion which was missed by NGS DNA and RNA sequencing within 5 days of tissue receipt. It should be noted that RNA sequencing is usually ordered after EPIC array and NGS DNA sequencing results are received.

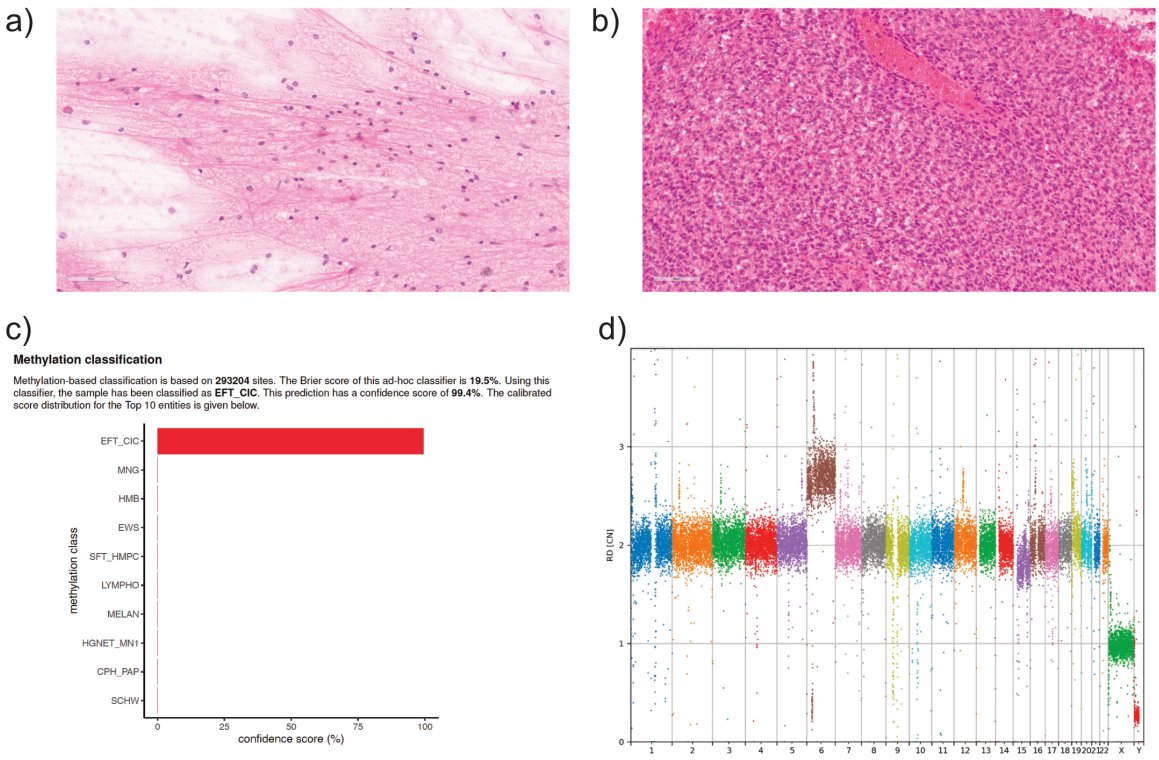


Figure 3-33 Accurate diagnosis of a CIC altered Ewing sarcoma by Rapid-CNS^2, which was described as a glioma by histology

Considering the aggressive nature of these tumours and considerably different treatment course as compared to the initial histology diagnosis of glioma, a swiftly established integrated diagnosis was essential.

3.3.17 Cost effectiveness

The cost of molecular analysis is a major barrier in its adoption across centres. I analysed the costs incurred for running Rapid-CNS^2 in different conditions. Flowcell costs vary depending on the pack purchased. Recommended shelf life of flowcells is 3 months. All prices are as per the ONT website store on 22nd September 2024. A comparison of Rapid-CNS^2 v1 and v2 is shown in **Figure 3-34**.

Rapid-CNS^2 v1: We used MinION flow cells for v1. Estimating two runs with two flow cells per week and reloading them twice, we were able to utilise a pack of 24 flow cells within three months. Each flow cell cost 475 EUR. The library preparation kit, which included three

reactions, was priced at 285 EUR. Flow cell washing incurred a cost of 32 EUR, and the gTube was 350 EUR. In total, the costs amounted to approximately 1,150 EUR per sample.

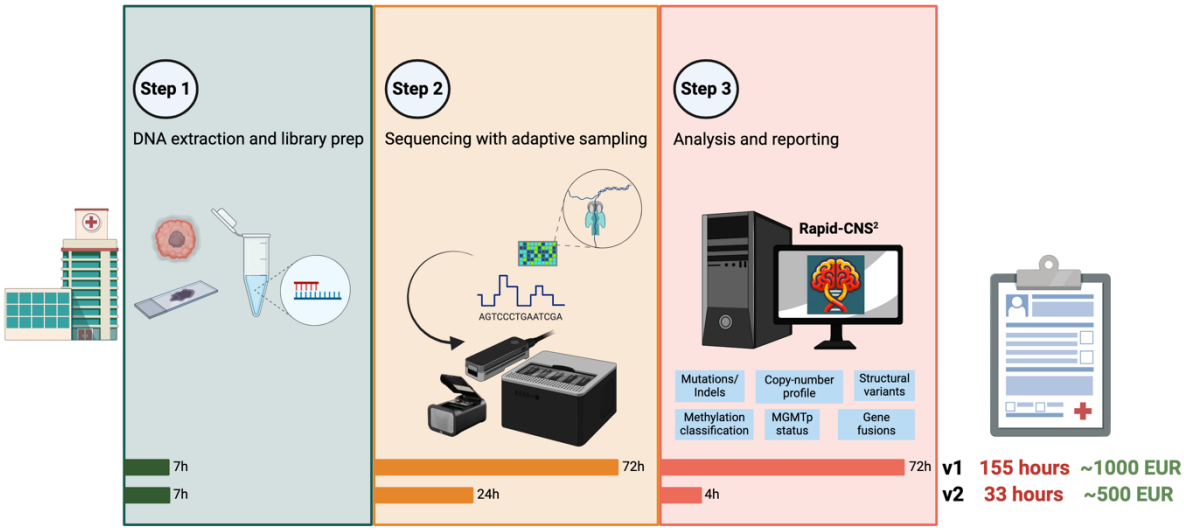


Figure 3-34 Comparison of Rapid-CNS^2 v1 and v2. (Costs for v1 and v2 for each sample using a MinION and PromethION flowcell respectively)

Rapid-CNS2 v^2: We used either MinION or PromethION flowcells for v2. Since we flush and reload the flowcells, we were able to use one flowcell for two samples. MinION flowcells cost 475 EUR each, approximating 240 EUR per sample. Library preparation with 1 reaction costs 95 EUR per sample. Flowcell washing costs 16 EUR and gTube costs 350 EUR, bringing the per sample cost to 700 EUR per sample. Using a PromethION flowcell that costs 390 EUR per sample, the workflow costs 850 EUR per sample. With the combination of the ultra-long kit and rapid kit, we could forego the gTube to reduce the cost to 350 EUR per sample with a MinION flowcell and 500 EUR per sample with the PromethION flowcell.

Thus, the improved configuration of v2 reduced costs and turnaround time while improving the accuracy and coverage even with the use of the more expensive PromethION device and flowcells.

3.3.18 Turnaround time

We achieved an average turnaround time of just 2 days from tissue receipt to the final report, which includes methylation classification, copy number profile, mutations, and structural variants for diagnostic samples. This is a significant improvement compared to the average

20 days turnaround time for the conventional workflow. When accounting for avoidable logistical and organisational delays, the entire pipeline from tissue to report took only 40 hours. For intraoperatively sequenced samples, it required less than 30 hours.

3.3.19 Code and data availability

Rapid-CNS^2 is available as a Nextflow pipeline that can be easily deployed with a single command and only requires a basic knowledge of command-line programming (https://github.com/areebapatel/Rapid-CNS2_nf). Code to run Rapid-CNS^2 v1 on the DKFZ cluster is available at https://github.com/areebapatel/Rapid-CNS2/tree/dkfz and to run Rapid-CNS^2 on the local workstation at Department of Neuropathology is at https://github.com/areebapatel/Rapid-CNS2_sh. All raw and processed data from this section is stored on the ODCF cluster.

3.4 Discussion

This chapter takes a journey through the establishment, validation, routine use and continued development of the Rapid-CNS^2 pipeline for the molecular characterisation of central nervous system (CNS) tumours. The modifications and optimisations introduced have led to notable improvements in efficiency, cost-effectiveness, and diagnostic accuracy, presenting promising implications for both clinical and research settings.

Nanopore sequencing has revolutionised molecular research with its portable devices, ease of use, native DNA/RNA sequencing, long reads and detection of base modifications, all at a relatively low cost. Recent studies have showed the potential of rapid whole-genome nanopore sequencing for methylation classification, but Rapid-CNS^2 takes this a step further[118,161-164,182]. By incorporating adaptive sampling, it targets clinically and prognostically relevant genomic regions while simultaneously covering sufficient CpG sites for methylation classification and fine copy number profiling. The demonstrated ability to report both methylation profiles and copy number variants (CNVs) within a surgical timeframe offers a significant advantage in CNS tumour diagnostics, particularly for intraoperative decision-making.

A key advantage of Rapid-CNS^2 is its ability to provide all molecular results simultaneously, which represents an improvement over the conventional diagnostic process, where tests are

typically ordered sequentially based on prior results. The immediate availability of comprehensive data proved particularly valuable in resolving cases where methylation classification was inaccurate. For instance, glioblastoma samples that were classified as inflammatory microenvironment could be correctly diagnosed by identifying the hallmark 7/10 signature in their copy number profiles. Additionally, for samples that did not belong to any classes represented in the classifier, the integration of mutational, copy number, and fusion data provided critical insights that helped refine the diagnosis. Importantly, in the few cases where potentially misleading diagnoses were issued, the results across multiple molecular layers did not support a clear or convincing conclusion. This highlights the procedural safety of the multi-modal approach, as no single erroneous result could unduly influence the final diagnosis. A recent study demonstrated that inaccurate AI model predictions can negatively impact the diagnostic accuracy of human clinicians, especially those with less experience[183]. While this study focused on radiology models, it is reasonable to assume that similar risks exist in neuropathology. Therefore, the multi-modal nature of Rapid-CNS^2 mitigates these risks by offering multiple layers of molecular data. Even if one data type is incorrect, the presence of other molecular results helps reduce the likelihood of a misdiagnosis by providing additional context for the neuropathologist.

This work demonstrated successful application of real-time methylation and copy number analysis for samples in an intraoperative timeframe at two independent centres. Previous studies demonstrating successful methylation classification using shallow whole genome sequencing were unable to achieve fine resolution for copy number variations in an intraoperative time frame[163,164,182]. Ad-hoc methylation classification performed reliably for all classifiable samples and those that could not be classified were explainable by tissue or clinical specificities. In cases where classification was not possible, accurate copy number profiles were obtained as early as 10 minutes into sequencing, providing crucial diagnostic insights. Two samples fell into categories that were not represented in the MNP v11 classification scheme, and therefore, could not be classified. This highlights the need for the development of MNP v12-based classifiers that can work with sparse intraoperative data to resolve such cases effectively. Although the approach yielded promising results, the relatively small sample size limits the ability to draw concrete conclusions about its broader clinical utility. Therefore, validating these findings in a larger cohort is essential to ensure robust and reliable application in clinical settings. To further enhance prediction confidence, a potential solution would be to use multiple published classifiers simultaneously. This approach has been

successfully tested by our collaborators in Nottingham, thus providing a comprehensive framework for classification[146].

The use of adaptive sampling greatly simplifies the targeted sequencing process, as it only requires a text file to assign regions of interest. This flexibility eliminates the need for complex primer designs and custom consumables, streamlining the workflow. By reducing the length of targeted panels, the pipeline can further enhance coverage. Recent developments have showcased the ability to barcode samples for multiplexing to make the process even more efficient[123]. A new framework BOSS-RUNS, allows for dynamic stopping of region targeting once sufficient coverage is achieved, further ensuring uniform coverage across regions of interest. These advancements can be incorporated into future iterations of Rapid-CNS[2], improving data quality while simultaneously reducing costs.

The ease of modifying targeted regions by simply editing a text file allows us to quickly adapt and customise the focus without re-designing and ordering custom panels as is required for NGS panel sequencing. One particularly valuable aspect is the flexibility to make these changes in real-time, even during the sequencing run. As initial evidence accumulates, the pipeline can be re-directed towards genomic loci that are of greater diagnostic and clinical relevance. In the pre-print by Deacon et. al 2024, this was shown for an intracranial schwannoma, where the schwannoma was detected within minutes[146]. The pathologist suspected a rare type of CNS schwannoma with a *VGLL* fusion as described before[184]. *VGLL1* and *VGLL3* were not a part of the gene panel that was being targeted initially, so the run was paused to add the coordinates of the gene to the bed file. Within an hour of targeting, the authors reported detection of split reads aligning to *VGLL3* and *CHD7* indicating a fusion. This shows the dynamic adaptability of the workflow. Methylation classification can considerably narrow down the variants relevant to a particular diagnosis. A useful feature for future work would be the ability to dynamically switch panels as soon as a confident methylation class is predicted. The workflow could begin with a broad CNS tumour panel with real-time methylation classification to generate a preliminary classification. Once a reliable classification is established, the panel could be refined to focus on regions of interest specific to the tumour type. For example, if the tumour is classified as CNS lymphoma, it warrants identification of variants in genes like *MYD88*, *CD79*, *PIM1*, *BTG2*, *PRDM1*, *BCL2/6*, *CARD11* etc.- genes not commonly associated with other CNS tumours[185]. Switching to a lymphoma specific panel would additionally enable detection of structural variants, such as translocations in *BCL2/6* that are not presently covered in the CNS tumour panel[186,187]. Detecting these structural

variants is crucial for accurate diagnosis and therapeutic decision-making, as they play a significant role in the pathogenesis and prognosis of CNS lymphomas. Additionally, long reads are capable of comprehensively profiling short tandem repeat (STR) regions relevant in neurological conditions like repeat expansion disorders[188-191]. A recent study targeted STRs with adaptive sampling and performed haplotype resolved assembly and DNA methylation profiling of their sites[190]. This could be incorporated as an additional feature in the Rapid-CNS^2 panels. Telomere length is known to play a significant role in CNS tumours particularly in glioblastomas[192-194]. Recent studies have showed that long-read sequencing is able to resolve allele-specific chromosomal telomere length[195,196]. This has the potential to observe an unprecedented detail in telomere biology and its effect on tumourigenesis and progression. Adding telomere regions to the panel could not only inform diagnoses but also add to our understanding of the disease biology.

One of the key results of this work was enabling use of small biopsy samples and the reduction of the minimum DNA input requirement to 500 ng, making the pipeline more accessible for cases with limited tissue samples. This is particularly beneficial in clinical settings where tumour samples may be scarce, such as in stereotactic biopsies or delicate surgical procedures. Additionally, the reduction in sequencing time to 24 hours not only expedites the diagnostic process but also allows for the reuse of flowcells, contributing to significant cost reductions. This makes Rapid-CNS^2 an affordable option for cases with limited tissue that is not sufficient for other molecular assays even in settings with limited resources, aligning with the broader objective of expanding access to advanced molecular diagnostics.

The detection of pathogenic single nucleotide variants (SNVs) and structural variants (SVs) remains crucial for comprehensive molecular diagnosis. While the findings in this work show that next-day pathognomonic SNV and SV detection is possible, I did not perform a systematic analysis of the types of errors incurred. A comprehensive analysis of the false positives and negatives would better inform if types of variants are better detected using Nanopore and which are more likely to be missed. Importantly, it should be recognised that NGS panel sequencing provides much higher coverage of target regions even to the order of 1000X. Using a single PromethION flowcell, the maximum on-target coverage we could achieve was around 60X. However, for somatic variant detection, a minimum coverage of 250X or more is typically recommended, and even higher coverage is suggested for detecting low-frequency variants[197]. I suggest that while nanopore-based SNV and SV detection can be a valuable additional tool in providing an integrated diagnosis, it should currently be viewed as a

complement to, and a reason for, confirmatory panel-based sequencing. At this stage, it is not yet suitable for novel variant discovery, particularly given the high false positive rate. However, with anticipated advancements in library preparation protocols, adaptive sampling methods, and flow cell architecture, further improvements in targeting may make this method more viable in the future.

A key limitation in this work was the inability to re-basecall and re-analyse all samples using the latest tools currently available. The samples processed with R9 flow cells could not be re-basecalled using newer models and tools developed for R10 chemistry. These newer tools are tailored specifically to the characteristics of the R10 chemistry and the newly introduced file types and formats, making them incompatible with data generated by R9 flow cells. Additionally, each library sequenced on a MinION flow cell demands a substantial amount of storage—between 200 and 500 GB per run—for basecalling and analysis and even more for PromethION flowcells. Beyond the significant storage requirements, each run also requires at least 3 hours of GPU time, coupled with CPU processing, to complete the analysis. Given these constraints on storage, compute resources, and no backward compatibility, I chose to use the data as it was originally reported during the sequencing and initial processing stages. The results presented in this work therefore reflect the pipeline configurations that were in place at the time of sequencing.

To address these limitations and provide comprehensive insights into the real-world accuracy of nanopore sequencing in clinical settings, a systematic study could be undertaken to evaluate the performance of updated basecalling and downstream analysis tools, specifically using clinical samples with varying tumour purities. While cell lines and well-characterised samples are typically used in benchmarking studies, they may not fully represent the complexity of clinical samples, particularly in terms of tumour heterogeneity, sample quality, and purity[198,199]. Therefore, using clinical samples with different known tumour purities would provide a more clinically relevant assessment of the performance of the platform. Additionally, a side-by-side comparison of the performance of older versus newer basecalling algorithms across the same dataset would help quantify any biases introduced by legacy tools.

I recognise that while nanopore sequencing offers significant benefits, including lower startup costs and a compact laboratory footprint, its accessibility to neurosurgery departments worldwide is still hindered by various financial and non-financial constraints. The non-financial barriers often include regulatory red tape, legal challenges related to the procurement and

supply of nanopore devices in certain countries, and an underdeveloped supply chain network, given that the company behind this technology is still in its growth phase. Despite these hurdles, the relatively low initial investment of around €50,000—compared to the approximately €500,000 required for Illumina-based sequencing—makes it reasonable to expect that nanopore technology could be adopted more widely as these logistical issues are addressed and the market matures. As described in the review by MacKenzie et. al 2023, PacBio- the other large company for long-read sequencing, has the Sequel II platform which achieves maximum read lengths exceeding 200 kb with accuracy of 87–92% and has an estimated cost per Gb ranging from USD 43 to 86 per Gb[200]. In comparison, Oxford Nanopore Technologies' PromethION platform, capable of generating reads over 1000 kb with a read accuracy of 87–98%, has a much lower estimated cost of USD 21 to 42 per Gb which is constantly decreasing. Additionally, its smaller space requirements make nanopore sequencing more feasible for institutions that might lack the infrastructure needed for more traditional sequencing platforms.

One of the significant challenges posed by nanopore sequencing, however, is the sheer volume of raw data generated. A single MinION flow cell can produce around 200 GB of raw FAST5 files, while a PromethION flow cell can yield over 500 GB per sample. Storing such vast amounts of data for extended periods is not only costly but also logistically prohibitive for many institutions. Furthermore, it is essential to preserve this raw data because basecalling technology is constantly evolving. As basecalling algorithms improve, reanalysis of raw data can yield more accurate results, making long-term data retention critical for both clinical and research purposes.

Another challenge is the computational expense associated with basecalling. Nanopore basecalling employs neural networks, and the most recent developments utilise transformers, which demand high-performance GPUs to operate efficiently. This computational requirement adds another layer of cost and complexity, as institutions must invest in advanced hardware to ensure rapid and accurate processing of sequencing data. These challenges need to be addressed to truly realise the benefits of nanopore sequencing—its portability, scalability, and lower initial costs making it a promising technology for broader adoption in the future, particularly as computational and data storage solutions continue to advance. This adds complexity and cost, but also presents a key opportunity for genomics facilities to modernise their infrastructure. Traditionally, genomics workflows have relied on multi-threaded CPUs in large high-performance computing environments, but platforms like NVIDIA Parabricks have

shown that incorporating multi-GPU systems can reduce processing times to under an hour[201,202]. Cancer Research UK's TRACERx EVO project, the latest expansion of the world's largest long-term lung cancer research program, exemplifies the potential of GPUs[203]. Early results from the Francis Crick Institute indicate that end-to-end analysis of whole human genomes can now be completed in just over two hours using NVIDIA Parabricks on GPUs, compared to the 13 hours previously required. This advancement, when extended across the workload of TRACERx EVO, is expected to save nearly nine years of bioinformatics processing time[204]. In this work, I used Parabricks to run DeepVariant, which significantly accelerated the variant calling process. With improved targeting and accumulation of more sequencing data, the computational demand will only increase, making it imperative to use GPU-accelerated platforms. Furthermore, Parabricks offers the ability to perform structural variant calling through de novo assembly, another computationally intensive task that benefits greatly from GPU acceleration[205]. This would allow more efficient detection of long and complex genomic aberrations in a reference-independent manner.

I developed a Nextflow pipeline for Rapid-CNS^2 to maximise its applicability across a wide range of environments[206]. Nextflow offers flexibility enabling the pipeline to be easily implemented on different systems, whether in local setups, high-performance computing clusters, or cloud platforms. This adaptability allows for broader deployment in both research and clinical settings, regardless of available infrastructure. Additionally, the use of containerisation, via Docker or Singularity, ensures that the pipeline remains reproducible and consistent, a key factor when working across different labs or institutions. Of note is that ONT's EPI2ME platform supports execution of Nextflow workflows through the MinKNOW UI. This means that Rapid-CNS^2 can be seamlessly imported and run directly through MinKNOW as a user-friendly workflow without the need for bioinformatics expertise.

Newer editions of the MinION are with integrated compute called MinION Compute or the latest announced edition MinION Mk1D that can be connected to modern laptops as well as the new iPad Pro with USB-C connectivity. In early 2024, ONT announced a new smaller device called the SmidgION with a new flow cell and sequencing device configuration. On the other hand, at London Calling 2024 they also introduced the large integrated sample to answer sequencing device called the ElysION- a standalone device with integrated MinION Mk1D or P2 Solo sequencer and onboard compute. Moreover, they also announced the TraxION, an all-in-one compact device that has a pipette-free approach and only requires an unprocessed sample to be loaded. Approximately the size of a MinION, it is claimed to have built-in nucleic

acid extraction, library preparation and integrated flowcell. This innovation has the potential to significantly reduce the need for dedicated wet-lab personnel and could be particularly advantageous for intraoperative sequencing, where homogenised tissue samples could be directly loaded into the system. Such advancements may facilitate the broader adoption of nanopore sequencing technology in smaller medical centres and clinics, particularly in point-of-care settings. Beyond hardware developments, ONT is actively pursuing enhancements to sequencing accuracy. One promising advancement is duplex sequencing, where one strand of DNA helps to correct errors in the complementary strand. Basecalling is also constantly improving with new transformer-based models combined with error correction approaches like HERRO[207]. ONT's base modification models now detect six different types of DNA and RNA modifications, with further developments in the pipeline. In light of these dynamic enhancements, it is reasonable to expect that nanopore sequencing will become even more accurate, user-friendly, and accessible in the coming years. Within the next 5-10 years, we may see this technology not only dominate fields like cancer diagnostics but also expand into a wide array of applications, from infectious disease monitoring and antimicrobial resistance detection to personalised medicine and environmental genomics. The portability, real-time capabilities, and continued reduction in cost are likely to drive wider adoption across both research and clinical settings globally. Furthermore, with the commitment of the long read sequencing to continuous improvement, the barriers for entry will likely lower, enabling smaller labs, field-based researchers, and even remote clinics to harness the power of real-time, high-throughput sequencing.

Nanopore sequencing holds transformative potential far beyond traditional diagnostics. The ability to generate long reads provides the unique advantage of resolving complex structural variants—such as large insertions, deletions, and rearrangements—that short-read sequencing technologies often miss. This capacity to uncover the full spectrum of genetic variation offers a more complete view of the genome, particularly in regions that are difficult to sequence, such as repetitive elements and highly homologous regions particularly telomeric regions. Moreover, long-read sequencing enables simultaneous detection of DNA methylation and other modifications on the same read as the genetic variations. This capability allows us to interrogate both genetic and epigenetic alterations within the same molecule, revealing insights into how these two layers of molecular information may interact. For instance, it could shed light on how epigenetic changes, such as methylation patterns, influence gene expression in the context of structural variants or mutations. This comprehensive view is

invaluable to understand CNS tumours where both genomic and epigenomic changes have been known to play critical roles.

In this context, Rapid-CNS^2 is not only pivotal for CNS tumour diagnostics but also offers a powerful resource for multi-omics research. By facilitating the simultaneous study of genome and epigenome alterations in routine clinical samples, data generated by Rapid-CNS^2 can help to unravel the interplay between different molecular processes, providing deeper insights into disease mechanisms, progression, and treatment responses. As the field of multi-omics expands, the integration of long-read sequencing technologies into research workflows will be key to advancing personalised medicine, identifying novel biomarkers, and unlocking new therapeutic targets.

In conclusion, I have developed a novel diagnostic workflow capable of providing reliable methylation-based tumour classification and copy number profiles within a 2-hour intraoperative timeframe, and delivering a comprehensive molecular profile, including the identification of MGMT promoter status, structural variants (SVs) and single nucleotide variants (SNVs), within 24 hours. This workflow could significantly transform the current standard of care for brain tumour diagnosis, streamlining the process and enhancing the depth of molecular insights available within a clinically relevant timeframe. Additionally, the dual genomic and epigenomic data generated by the workflow would serve as a robust resource to investigate the nuances of CNS tumour subtypes. However, for this advancement to be fully realised in practice, it is essential that healthcare professionals receive proper training to interpret and analyse the complex data produced by these assays. The role of the neuropathologist is changing, with increasing responsibilities to interpret complex molecular data in real-time. They will be expected to integrate these new molecular findings with traditional diagnostic tools, such as histopathology, clinical and radiological data, to provide comprehensive diagnoses. It is only through synthesizing these various sources of information that we can achieve better patient outcomes and enable the delivery of truly personalised therapies. I hope that this diagnostic tool represents a pivotal step forward in the treatment of brain tumours. As we continue to refine this technology and expand its adoption, it holds great promise for improving both the accuracy and speed of diagnosis, thereby facilitating more tailored and effective treatment strategies for patients.

Chapter 4 MNP-Flex

4.1 Introduction

Traditional methods for DNA methylation profiling typically rely on generating methylation profiles through methylation arrays, followed by supervised classification trained on a comprehensively annotated reference dataset. This has now become a standard diagnostic approach in adult and paediatric neuro-oncology, with methylation profiling being recommended by the WHO guidelines[72,73,208]. However, as technology evolves, a range of methods for interrogating the methylome at different resolutions and target regions has emerged. Whole-genome bisulphite sequencing (WGBS) is considered the gold standard, offering the most comprehensive and high-resolution methylation maps at the single-base level. Despite its accuracy and comprehensiveness, WGBS is expensive, requires significant quantities of input DNA, and often results in reads that lack usable methylation information. Alternative methods, such as methylation panel sequencing employ restriction enzymes or hybridisation capture to reduce costs by focusing on specific genomic regions. Microarray-based technologies like the Infinium HumanMethylation450 (450K) and MethylationEPIC (850K) arrays are extensively used to report specific CpG sites across the genome[209,210]. More recently, third-generation sequencing techniques, such as nanopore sequencing, have allowed methylation data to be obtained directly from native DNA, without the need for chemical conversion. The previous section of this work, along with other studies, has demonstrated the utility of low-coverage stochastic nanopore sequencing for rapid, affordable DNA methylation-based classification of CNS tumours[112-114,148,211-213]. However, this approach typically provides binary methylation information from a random subset of CpG sites, rather than the more detailed beta values common in array data.

These diverse technologies have each demonstrated high concordance in results, but their different coverage depths, errors owing to sequencing and specific genomic target regions have necessitated platform-specific classification approaches. Machine learning has been extensively applied in this space, with random forest (RF) models being a popular choice for

array-based methylation classification. However, RF models are limited by fixed feature sets and are often platform-specific, reducing their applicability across different methylation profiling techniques. In the previous section, I proposed an ad-hoc RF approach to apply the classification scheme to low-coverage nanopore sequencing, but this required training a new model for each sample. This endeavour was computationally expensive, time-intensive, and made cross-sample comparison difficult[114,148,211-213]. More recently, neural network-based models were proposed to handle sparse methylation data, providing more robust predictions for brain tumour classification[112,113]. However, a precise and flexible model that can handle data from multiple platforms while maintaining high prediction accuracy is still urgently needed.

The application of machine learning to tabular data has gained significant traction in recent years, with several popular approaches emerging as go-to solutions. Among these, random forests, gradient boosting, and neural networks are widely recognised for their effectiveness in a variety of tasks. Each model has its unique strengths and weaknesses, and the choice of which to use often depends on the nature of the data and the specific problem being addressed. In this section, I will explore how these models perform in the context of tabular data, drawing on both theoretical insights and practical evidence.

4.1.1 Random Forests

First introduced by Breiman et. al 2001, random forests have become a staple in the machine learning toolkit[214]. At its core, the random forest algorithm builds multiple decision trees through bagging—short for Bootstrap Aggregating. Essentially, the model creates several different trees, each trained on a random subset of the data, and makes predictions by averaging the outputs of these trees. This approach is particularly effective at reducing overfitting, as it ensures that no single tree dominates the decisions made by the model. One of the major advantages of random forests is their robustness. They can handle noisy data well and often perform reliably even in situations where the data contains irrelevant or redundant features[215]. However, while random forest tend to perform well in many tasks, they require fixed data points and cannot deal with missingness in data. Additionally, although they provide some level of interpretability through feature importance scores, the sheer number of trees can make it challenging to fully understand the model decisions[216]. Random forests have been used to classify CNS tumours in all available versions of the MNP classifier[62,63]. When

Illumina changed their EPIC array to v2, this necessitated training a new classifier with sites overlapping the v2 array since some sites from v1 had been made redundant[210].

4.1.2 XGBoost

XGBoost, which stands for Extreme Gradient Boosting, takes a different approach by using boosting rather than bagging. Boosting is a sequential process; each new tree is built to rectify the mistakes in the previous one. This strategy allows XGBoost to refine its predictions iteratively, leading to high accuracy, particularly in tasks involving tabular data[217]. XGBoost also incorporates advanced techniques like L1 and L2 regularisation, helping to prevent overfitting, a common problem in machine learning. One of the reasons XGBoost has become so popular is its ability to handle a variety of data challenges, such as missing values and categorical variables, with minimal preprocessing[218]. It is also highly efficient, thanks to its optimisation strategies, making it a preferred choice in competitive machine learning. However, the sequential nature of XGBoost also makes it more difficult to interpret compared to simpler models like Random Forests. The intricate way it builds and combines trees means that understanding exactly how features contribute to the final prediction can be quite complex[219]. Nevertheless, approaches such as Shapley values have been used reliably to add layers of explainability to XGBoost predictions.

4.1.3 Neural Networks

Neural Networks, particularly deep learning architectures, have garnered significant attention in recent years due to their success in fields like image recognition and natural language processing[220]. However, neural networks often struggle to outperform tree-based models like Random Forests and XGBoost for tabular data. This is largely because tabular data typically consists of fewer features and less complexity compared to unstructured data like images or text, where neural networks excel. For smaller datasets, neural networks are often prone to overfitting unless measures like dropout or batch normalisation are used[221]. Additionally, neural networks typically require careful feature scaling and preprocessing, whereas tree-based models like Random Forests and XGBoost handle such tasks natively. The lack of interpretability is another significant drawback. While neural networks are excellent at capturing non-linear relationships, they are often described as “black-box” models, since it is difficult to explain how the model arrived at a particular prediction[222].

4.1.4 Comparison for Tabular Data

For most tasks involving tabular data, tree-based models like random forests and XGBoost tend to outperform neural networks[223,224]. This is especially true when the dataset is small or the relationships between features are not highly complex. In practical applications, XGBoost often leads to higher predictive accuracy than both random forests and neural networks, owing to its ability to minimise both bias and variance[225]. That said, random forests are still an excellent choice in many cases, particularly when interpretability and ease of use are important considerations. They also tend to train faster and require less fine-tuning compared to XGBoost. In contrast, while neural networks shine in tasks involving high-dimensional, unstructured data, their performance on tabular datasets tends to lag behind. Neural networks typically require larger datasets to train effectively and, even then, often need significant tuning to achieve comparable results to tree-based methods[223]. They also require considerable computational resources, which can be a limiting factor, particularly in situations where resources are constrained. In view of the nature of sequencing based methylation data, XGBoost provided the best balance between error and missing value handling and data as well as computational requirements.

Additionally, the previous approaches have been trained on a 91 class scheme of classification based on the Capper et. al dataset. The current version v12.8 (as of October 2024) encompasses 184 methylation subclasses, thus doubling the number of identifiable classes. The newly identified classes often constitute rare subtypes that could be differentiated owing to the unsupervised clustering of the data followed by thorough investigation of additional molecular layers. These classes are now recommended to be identified by methylation profiling in the WHO classification of CNS tumours 2021[73].

In this work, I propose MNP-Flex - a v12 compatible XGBoost-based model as a robust, scalable alternative to neural networks for DNA methylation classification. XGBoost, a powerful gradient-boosting algorithm is able to handle sparse data, feature variability, and large datasets efficiently. The XGBoost-based framework in this work is trained on a fixed reference dataset and designed to classify methylation data generated by diverse platforms, including WGBS, targeted methyl-seq, nanopore whole-genome as well as adaptive sampling-based sequencing, and microarray platforms (e.g., Illumina 450K, EPIC, EPICv2). This approach offers several key advantages over previous models, such as faster training times, reduced computational costs, and improved generalisation across different methylation

platforms. Additionally, the inherent scalability of the model allows for rapid updates and re-training as new reference datasets become available, ensuring that it remains compatible with emerging CNS tumour methylation atlases.

By employing XGBoost, I aim to create a unified framework that can accommodate the growing diversity of DNA methylation data while delivering precise, platform-agnostic tumour classification.

4.2 Methods

4.2.1 Model training

Martin Sill and I trained and evaluated the MNP-Flex model respectively. The classification model was constructed using gradient-boosted decision trees, specifically utilising the XGBoost algorithm (R package xgboost v2.01)[226], which is well-regarded for its efficiency and performance in handling large datasets and complex classification tasks. The training dataset, was the same as that used for the MNP classifier v12. The dataset comprised 7,495 biological samples distributed across 184 distinct methylation classes. These methylation classes are comprehensively described on the classifier website (https://www.molecularneuropathology.org) and in A.2.1. For model training and evaluation, the dataset was partitioned into two subsets: a training set consisting of 70% of the samples and a validation set comprising the remaining 30%. This split was carried out using the createDataPartition function in the caret package (R package caret v6.0-94)[227], ensuring that the distribution of methylation classes was preserved across both sets.

The following extract is adapted from Patel et. al 2024[117]:

“Preprocessing of the raw signal intensities was performed using the minfi Bioconductor package (version 1.21.459)[228]. Raw intensities were extracted from IDAT files generated from Illumina EPIC or 450k arrays as applicable. To merge the data from these two array platforms, the intersection of the CpG probes common to both platforms was selected, using the combineArrays function from the minfi package[228]. This step was crucial to ensure comparability and consistency across the different datasets. Background correction was applied to adjust the signal intensities, where the 5th percentile of negative control probe intensities was shifted to 0, reducing any non-biological noise in the data. Additionally, dye-

bias correction was performed by scaling the mean intensities of normalisation control probes to 10,000 for both color channels (red and green), ensuring that technical artifacts introduced during the scanning process were mitigated.

To further account for batch effects arising from the different types of sample material (FFPE vs. frozen tissue) and array platforms (450k vs. EPIC), we employed the removeBatchEffect function from the limma package (version 3.30.11). This function fits univariate linear models to the log2-transformed intensity values, allowing for the removal of systematic biases introduced by differences in tissue processing or platform technology. Importantly, both methylated and unmethylated signals were corrected individually, as the processing biases might affect each signal type differently.

Following batch correction, beta-values were computed from the normalised intensities. These beta-values represent the proportion of methylation at each CpG site and were calculated using an offset of 100, as recommended by Illumina, to prevent division by zero and to stabilise variance at low intensities. We then performed a stringent filtering of CpG probes. Probes were retained based on the filtering criteria proposed by Zhou et al. (2017)[229], which ensures that only reliable and biologically relevant probes are included. Specifically, probes located on the X and Y chromosomes were excluded to avoid sex-specific methylation bias, resulting in a final set of 357,521 probes. To reduce the dimensionality of the dataset and focus on the most informative methylation sites, we further filtered the dataset to retain only the 100,000 CpG probes with the highest standard deviation across all samples. This step ensured that the most variable and potentially discriminative probes were prioritised in the model training process.

In preparation for training the classification model, the pre-processed beta-values were binarised by applying a threshold of >0.6. This binarisation step was undertaken to facilitate the future application of the model to lower-coverage sequencing-based data sources, where methylation calls are typically dichotomised. By setting a threshold, we transformed the continuous beta-values into binary categories, representing methylated and unmethylated states.

The XGBoost model was trained using the using the “multi:softprob” objective function that aims to minimise the multiclass log loss, i.e. negative log-likelihood of a logistic model which is also known as cross-entropy. A loss function typically used in probabilistic forecasting models, i.e. same is used in the v12.8 MNP-RF glmnet calibration model. The training process

was conducted for 2,306 iterations, with a learning rate (eta) set at 0.01 to control the step size in the optimisation process. The training was monitored using early stopping, which halted the training process when no improvement in the multiclass log loss was observed for a pre-defined number of rounds. This strategy helps prevent overfitting and ensures that the model generalises well to unseen data."

4.2.2 Data collection

I collected samples from diverse sources as shown in **Table 4-1**. The data encompassed methylation profiles obtained from four different technologies – whole genome bisulphite sequencing (WGBS), methylation panels, Nanopore whole genome sequencing, and Rapid-CNS[2].

Table 4-1 Details of non-array samples used for validating MNP-Flex

Sequencing type	***No. of samples***	***Site / Project***	***With matched array***
WGBS	25	KiTZ / KickCan	25
	34	KiTZ / ICGC PBCA-DE	34
	21	KiTZ / internal	21
Methylation panels	11	University Hospital Basel	11
	15	Ghent University	12
	1	University of Cincinnati	1
Nanopore WGS	40	University Hospital Oslo	None
Rapid-CNS[2]	252	University Hospital Heidelberg	210
	49	University of Nottingham	44

4.2.3 Whole genome bisulphite sequencing

The following extract is adapted from Patel et. al 2024[117]:

"Whole genome bisulphite sequencing libraries were sequenced at the Hopp Children's Cancer Center (KiTZ), Heidelberg. Samples were prepared for the WGBS library using the "Swift Accel-NGS Methyl-Seq DNA" kit and sequenced on the Illumina HiSeq X Ten V2.5 in paired-end mode, with one lane per tumour sample, resulting in an average genome coverage

of ~30x per sample. WGBS sequencing data were analysed using methylCtools (https://github.com/hovestadt/methylCtools) as part of the ODCF Bisulfite core workflow (https://github.com/DKFZ-ODCF/AlignmentAndQCWorkflows; AlignmentAndQCWorkflows:1.2.73-2)[230,231]. In brief, methylCtools builds upon BWA and adds functionality for aligning bisulphite-treated DNA to a reference genome in a similar manner described previously33.Sequencing reads were adaptor-trimmed and translated to a fully C-to-T converted state. Alignments were performed against a single index of both in silico bisulphite-converted strands of the human reference genome (hs37d5 including PhiX) using BWA[121]. Previously translated bases were translated back to their original state, and reads mapping antisense to the respective reference strand were removed. Single-base-pair methylation ratios (β-values) were determined by quantifying evidence for methylated (unconverted) and unmethylated (converted) cytosines at all CpG positions. Only properly paired or singleton reads with mapping quality of ≥1 and bases with a Phred-scaled quality score of ≥20 were considered. I used processed WGBS data from the publicly available PBCA-DE cohort on the ICGC portal[230]."

4.2.4 Methylation panels

The following extract is adapted from Patel et. al 2024[117]:

"Twist targeted methylation sequencing was performed at Ghent University, University Hospital Basel and University of Cincinnati respectively. DNA was extracted from FFPE tissue. 200 ng DNA was used as input for the Twist Human Methylome Panel[232]. The protocol provided by the Twist Targeted Methylation Sequencing Protocol was followed. The libraries were sequenced on a Novaseq (2x150 cycles). Methylation values were extracted using a Nextflow pipeline (https://nf-co.re/methylseq/1.6.1)."

4.2.5 Nanopore whole genome sequencing

The following extract is adapted from Patel et. al 2024[117]:

"Nanopore whole genome sequencing was performed at the University Hospital Oslo using the following protocol: gDNA was extracted from fresh or fresh-frozen tumour biopsies with the Qiagen Blood & Tissue mini kit. Briefly, 10-30 mg of tissue were homogenised in ATL buffer in a Tissuelyzer bead mill at 30 Hz for 30 seconds, followed by digestion with proteinase

K for 3-16 hours. Buffer AL and RNAse were added to the sample and incubated at RT for 5 minutes, followed by incubation at 70 degrees C for 10 minutes. 100% EtOH was added to the sample prior to washing and elution on spin columns. DNA purity was evaluated with Nanodrop (260/230 > 1.8 and 260/280 > 1.9 was deemed sufficient) and concentration measured with Qbit DNA Broad-range kit.1-3 µg of gDNA were used as input for sequencing library preparation with Ligation Sequencing kit V14 (SQK-LSK114) according to manufacturer's protocol (Ligation sequencing DNA V14). 300 ng of DNA library were loaded onto PromethION flow-cells (FLO-PRO114M) on a P24 sequencing device, one library per flow-cell and sequenced for 80 hours. Flow-cells were washed and reloaded if necessary after 24 or 48 hours of sequencing (Flow-cell wash kit, EXP-WSH004). Live basecalling, methylation calling and mapping (hg38) was performed via the MinKNOW software (version 23.07) with Dorado (version 7.1.14). Basecalling was performed with the super-high accuracy model (dna_r10.4.1_e8.2_400bps_sup@v4.1.0), sequences below the quality threshold of 10 were excluded from further analysis. Per-site methylation extraction and across-strand aggregation from modified .bam files was performed in the epi2me-labs suite through the wf-human-variation (version 1.8.1) workflow with modkit (v0.2.0) or modbam2bed (v0.10.0). Whole-genome methylation .bed files were cross-referenced with EPIC probe genomic locations with the bedtools intersect function."

4.2.6 Data pre-processing

I formulated a BED file constituting all sites from the Illumina Infinium MethylationEPIC array and the *MGMT* promoter region (`chr10:129466536-129467536`, hg38 genome[233]). The file constituted columns called chr (chromosome), start (start position), end (end position), IlmnID (Illumina CpG ID). This will be called MNP-Flex_sites.bed for the rest of this work. In order to maintain uniformity, I lifted over data bedmethyl files to hg38[127]. I ran bedtools intersect with `-wa` and `-wb` options on all hg38 compliant bedmethyl files to obtain the sites that were represented in the MNP-Flex_sites.bed file. I converted them to files with columns called- chr, start, end, coverage, methylation_percentage and IlmnID. I required all input files to MNP-Flex have at minimum columns called- coverage, methylation_percentage and IlmnID.

4.2.7 Model execution

To execute the model, I developed an R script, which was run through a bash script for streamlined automation. Additionally, I formulated a Docker container to facilitate the end-to-end execution of the model in a consistent and reproducible environment. The container was also employed to build the backend of a publicly available website, mnp-flex.org, developed by Daniel Schrimpf. This website enables users to run the model interactively. Users can upload input files in the specified format and subsequently download a comprehensive report. The report includes the coverage distribution, methylation value distribution, predictions, and scores across all four levels of methylation classification (subclass, class, family, and superfamily), as well as the prediction of *MGMT* promoter status, if applicable.

4.2.8 Ground truth data

Methylation array data analysed using the MNP v12.8 model (hereafter referred to as MNP-RF) was considered as ground truth for samples with matched array data available. Array data was generated using either frozen or FFPE tissue as previously described[19,144]. For nanopore WGS and methylation panel samples without matched array data, I inferred ground truth based on recorded histopathological evaluation and/or pathognomonic molecular alterations as described in the sections below.

4.2.9 Concordance analysis

The following extract is adapted from Patel et. al 2024[117]:

"I subset data from all validation samples to sites present in the Illumina InfiniumEPIC Methylation array and the *MGMT* promoter region. For methylation array data, I compared subclass and family level predictions to the corresponding MNP-RF predictions. For non-methylation array samples, I calculated concordance for MNP-flex samples based on predictions made for corresponding EPIC array profiles by MNP-RF or available neuropathology data assessment. I calculated confidence intervals using the binom R-package. I generated plots using ggplot2, ggsankey, ggridges, patchwork, and related R packages for visualisation. I obtained MNP-flex scores from analysing samples gathered from FFPE and frozen sources by testing with non-parametric equivalence test available through the R-package TOSTER using an upper and lower equivalence bound of 0.01[234]"

4.3 Results

4.3.1 Model training

With the advances in sequencing-based approaches for methylation analysis, including but not limited to nanopore sequencing, WGBS, and methylation panels, I sought to extend the utility of the latest version of the MNP classifier (v12) to accommodate such data. The MNP v12 classifier employs a hierarchical structure comprising 184 subclasses, 143 classes, 75 families, and 34 superfamilies. Martin Sill and I developed and evaluated MNP-flex, a platform-agnostic CNS tumour methylation classifier. The model uses binarised values to specifically allow for low coverage methylation calls as input data and account for the technical differences across sequencing technologies employed for methylation calling. The model was trained on a binarised version of the reference dataset used to develop the array-based MNP-RF model.

Upon completion of training, the model achieved a multiclass log loss of 0.1969 on the validation dataset, indicating a strong predictive performance.

4.3.2 Thresholding

I investigated the effect of prediction score on accuracy. I observed that accuracy increased proportionally with prediction score. On analysing the array and non-array validation datasets, I determined that optimal thresholds for MNP-Flex classification are ≥0.7 for array-based samples and ≥0.3 for sequencing-based samples (A.2.3). When applying these thresholds, the model achieved an overall family-level accuracy of 99.6% and subclass-level accuracy of 99.2% across all platforms. These thresholds offer a balance between sensitivity and specificity, making MNP-Flex a reliable tool for methylation-based CNS tumour classification.

4.3.3 Validation on methylation array data

I tested the MNP-Flex model on the entire MNP dataset, consisting of over 90,000 samples that were uploaded to the molecularneuropathology.org website. These samples had been previously classified using the MNP-RF model. As the MNP-RF model considers scores ≥0.9 as reliably classified, I imposed a similar threshold to validate MNP-Flex[19,144]. However, literature suggests that the threshold could be lowered to 0.84 or even 0.7 to account for real-

world variability[235]. For this reason, I tested both a strict cut-off (≥0.9) and a more lenient one (≥0.7) for model validation. Using a threshold of 0.7 for MNP-RF, I compared 78,833 samples and achieved a subclass-level accuracy of 92.7% (95% CI: 92.5% to 92.8%) and a family-level accuracy of 95.7% (95% CI: 95.5% to 95.8%). The subclass predictions displayed an AUC = 0.887 (A.2.2).

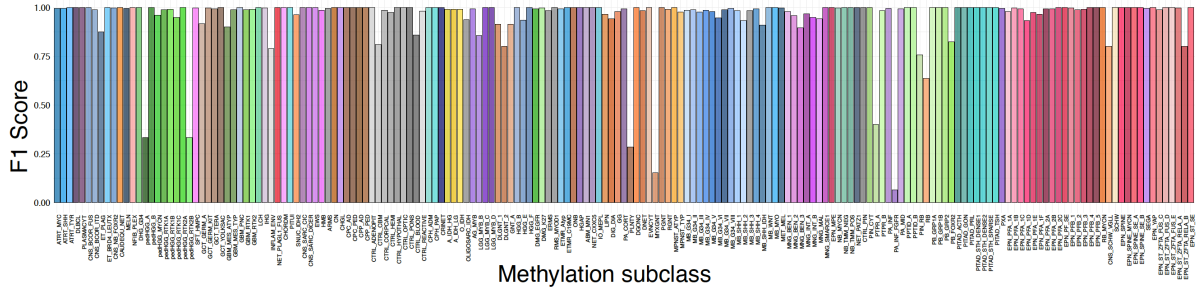


Figure 4-1 F1-scores for the methylation array dataset comprising 78,833 samples (Reprinted from Patel et. al 2024[117])

Further filtering samples for which the MNP-Flex subclass score was ≥0.7, the model achieved an accuracy of 98.5% (95% CI: 98.4% to 98.6%) for 58,410 samples across 182 subclasses. Majority of these subclasses (176 out of 182) had F1 scores greater than 0.5, with 163 subclasses having F1 scores ≥0.9, demonstrating the robustness of the model for subclass prediction (**Figure 4-1**).

4.3.4 Validation on non-array data

I evaluated MNP-Flex on sequencing-based data from 448 samples across four technologies: whole-genome bisulphite sequencing (WGBS), nanopore sequencing, methylation panels, and Rapid-CNS[2]. As shown in **Figure 4-2**, non-array validation data originated from seven institutions.

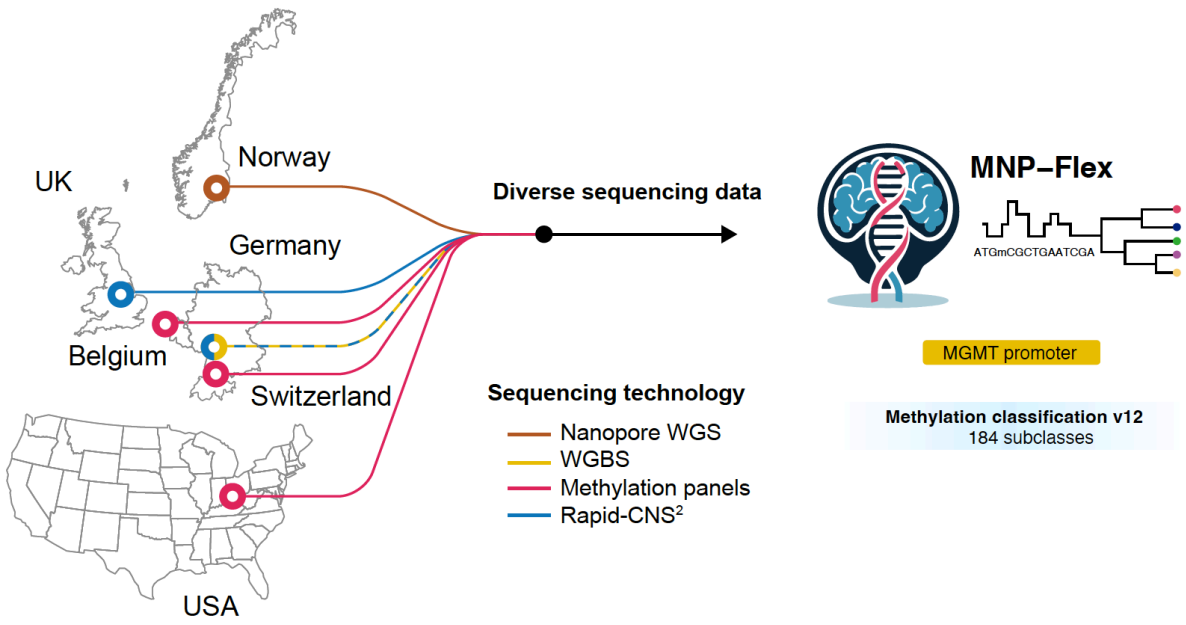


Figure 4-2 Sources of validation data for MNP-Flex (Reprinted from Patel et. al 2024[117])

For all non-array samples without thresholding, I observed a subclass accuracy of 65.9% (95% CI: 61.0% to 70.5%) and family accuracy of 91.9% (95% CI: 88.9% to 94.1%). However, applying a prediction threshold of 0.3 increased subclass accuracy to 82.8% (95% CI: 77.0% to 87.4%) and family accuracy to 99.5% (95% CI: 97.5% to 99.9%). Accuracy over different prediction scores for each technology are shown in A.2.3. Sankey plots for both conditions (prediction score ≥ 0.3 and ≤ 0.3) are displayed in **Figure 4-3** and **Figure 4-4** respectively. As indicated by the accuracy metrics, high confidence samples incur errors that tend to occur within their respective families. On the other hand, low confidence samples are more prone to cross-family predictions. Confusion within the same family would usually not have major consequences for patient diagnosis and treatment, while for instance, misprediction of a high grade tumour as low grade would be deleterious. This further supports the rationale for thresholding prediction scores to improve performance.

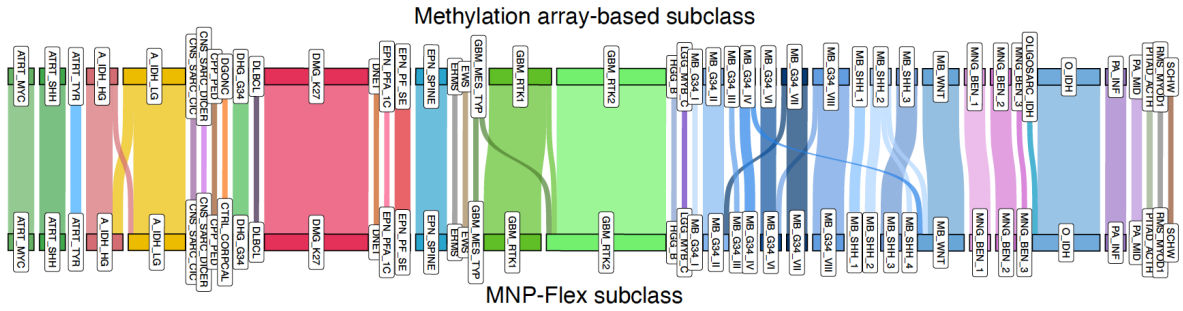


Figure 4-3 High confidence MNP-Flex predictions for non-array samples (score ≥ 0.3) (Reprinted from Patel et. al 2024[117])

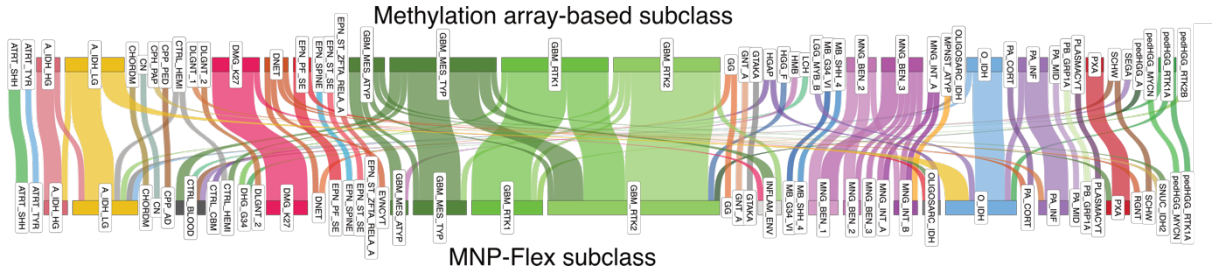


Figure 4-4 Low confidence MNP-Flex predictions for non-array samples (score ≤ 0.3) (Reprinted from Patel et. al 2024[117])

Since data was derived from four different non-array technologies, I delved deeper into the nuances of each technology. **Figure 4-5** presents a comparative analysis of the employed DNA methylation profiling methods, evaluating their accuracy, score distributions, and the percentage of missing sites. Barplots indicate accuracy with 95% confidence interval for samples processed using whole genome bisulphite sequencing (WGBS), Twist methylation panels, Rapid-CNS^2 in Nottingham and Heidelberg, Nanopore whole genome sequencing and the MNP methylation array dataset. The solid coloured bars indicate subclass level accuracy and bars with increased alpha indicate family level accuracy. Nanopore whole genome sequencing samples did not have matched array predictions, thus, family-level predictions were inferred from histological and molecular findings. Density plots indicate scores for subclass prediction while boxplots denote percentage of missing CpG sites in each dataset.

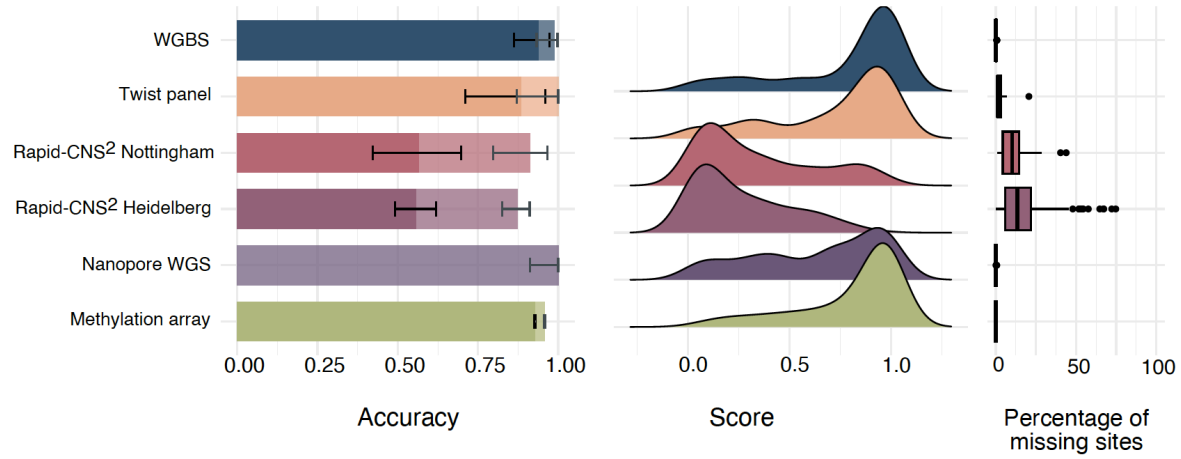


Figure 4-5 Comparison of MNP-Flex performance over different technologies without threshold (Reprinted from Patel et. al 2024[117])

As indicated in A-14, accuracy of subclass and family level predictions increased on thresholding all methylation array samples with scores ≥ 0.7 and non-array samples ≥ 0.3. Improved accuracies for each technology are demonstrated in **Figure 4-6**. Detailed description of individual results is given in the sections 0 - 4.3.4.4 below.

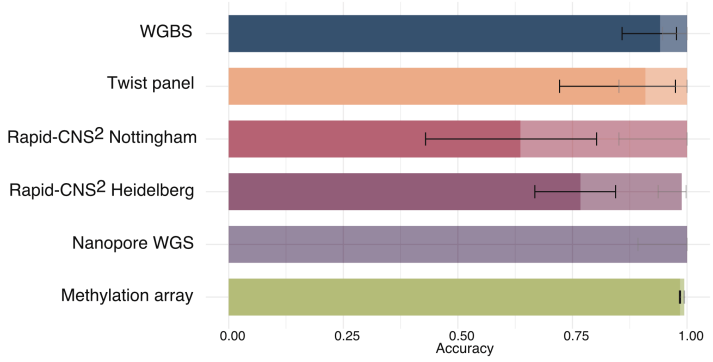


Figure 4-6 MNP-Flex accuracy over different technologies MNP-Flex accuracy for different technologies for array samples with scores ≥ 0.7 and non-array samples ≥ 0.3 (Reprinted from Patel et. al 2024[117])

4.3.4.1 Whole genome bisulphite sequencing

I accumulated 80 pre-processed WGBS bedmethyl files from publicly available repositories and internal projects as shown in **Table 4-1**. As indicated in **Figure 4-5** and **Figure 4-6**, MNP-Flex accurately predicted the methylation subclasses for 93.8% of WGBS samples (75 out of 80). Four misclassified samples were classified within the same methylation family, while one sample was predicted as “inflammatory microenvironment” which indicates inflammatory components of the true subclass. WGBS data had an average missingness of 0.13 % and

mean prediction score of 0.77 (**Figure 4-5** and **Figure 4-6**). While tools for WGBS analysis usually set a coverage threshold for methylation calculation, I did not apply a coverage filter to the data. Higher coverage data (>10X) like the publicly available ICGC PBCA-DE samples, achieved mean prediction scores of 0.92. I speculate that while selecting only high coverage sites does improve scores, the reduced number of sites would negatively affect model performance.

4.3.4.2 Nanopore whole genome sequencing

I received bedmethyl files for 40 brain tumour samples sequenced at University Hospital Oslo. The samples had been sequenced to an average depth of 30X over the whole genome. There was an average missingness of 0.02 % over all samples while they achieved a mean prediction score of 0.63 (**Figure 4-5**). Since these samples did not have matched methylation arrays as ground truth, I adopted a conservative approach and inferred "methylation families" from the reported molecular alterations as well as histological and clinical data (A.2.4). For example, presence of a H3K27 alteration confirmed the prediction of "diffuse midline glioma, H3K27-altered" or detected ZFTA fusion supported the prediction of methylation family "supratentorial ependymoma, ZFTA fusion-positive". Predicted methylation families were concordant in all nanopore whole genome sequencing samples, with inferred families aligning with corresponding molecular alterations.

4.3.4.3 Methylation panels

Methylation panels by Twist were used for the enzymatic EM-seq based approach. Martin Sill and I received 27 samples processed using Twist methylation panels from Ghent University, University of Cincinnati and University Hospital Basel. As demonstrated in **Figure 4-5** and **Figure 4-6**, the samples had an average missingness of 2.7 % but achieved mean prediction scores of 0.73. Two samples did not have matched array but were diagnosed as "alveolar rhabdomyosarcoma". Since this diagnosis exactly corresponds to one methylation class, I considered this as ground truth. Out of 27 samples sequenced using Twist methylation panels, 25 were correctly classified at the subclass level (A.2.5) One glioblastoma RTK1 subtype sample was classified as glioblastoma RTK2, while the other (Twist_3) had very few sequencing reads and a low prediction score ≤0.03. On excluding Twist_3, the average CpG missingness reduced to 1.9 % and average prediction score increased to 0.76.

4.3.4.4 Rapid-CNS^2

174 out of 194 (89.7%) classifiable samples with matched arrays were accurately assigned to the correct methylation family, with 128 samples (66.0%) correctly classified at the methylation subclass level (**Figure 4-5**). When excluding samples with confidence scores ≤0.3 (n=74), the accuracy at the methylation subclass level improved to 89.1%, while the accuracy at the family level rose to 98.7% (**Figure 4-6**). 11, 835 missing sites. Similarly, for the Nottingham Rapid-CNS^2 dataset, methylation subclasses were correctly predicted for 26 out of 41 samples (63.4%), and methylation families for 37 samples (90.2%) (**Figure 4-5**). After excluding samples with scores ≤0.3, the accuracy increased to 78% for methylation subclasses and 100% for methylation families (**Figure 4-6**).

The model struggled with Rapid-CNS^2 data due to a large number of missing methylation sites (16.56 %), leading to lower subclass accuracy. Reassuringly, MNP-Flex classified samples from this dataset with lower prediction scores (**Figure 4-5)**. The Heidelberg Rapid-CNS^2 dataset had an average of 17.59 % missing sites.

Table 4-2 MNP-Flex results for Rapid-CNS^2 Heidelberg dataset with matched arrays

Device	*Flowcell*	*Sequencing time*	*No. of samples*	*Missing sites (%)*	*Average score*	*Accuracy (%) Family / subclass*
MinION	R9	72	44	5.76	0.35	95.45 / 84.09
GridION	R9	24	18	27.31	0.16	83.33 / 55.55
		48	11	34.55	0.14	81.82 / 36.36
		72	48	14.51	0.17	83.33 / 62.5
GridION	R10	24	55	21.43	0.28	87.27 / 63.63
P2 Solo	R10	24	18	2.99	0.49	88.88 / 66.66

As indicated in the meta-information for the dataset shown in **Table 4-2**, I observed that samples sequenced on the MinION with R9 flowcells for 72h had highest accuracy. While this seems counterintuitive, all samples sequenced on the MinION were archival and were carefully selected for high tumour purity to facilitate set-up of the workflow. The flowcells were re-loaded twice during the run. Additionally, samples belonging to only 4 major glioma types namely, glioblastoma, oligodendroglioma, astrocytoma and pilocytic astrocytoma were selected for the set-up phase. This led to highly optimised results that would be difficult to

achieve in a real world setting with limited tissue and personnel hours available. As expected, samples run on the P2 Solo with R10 flowcells had the lowest number of missing values and highest average score.

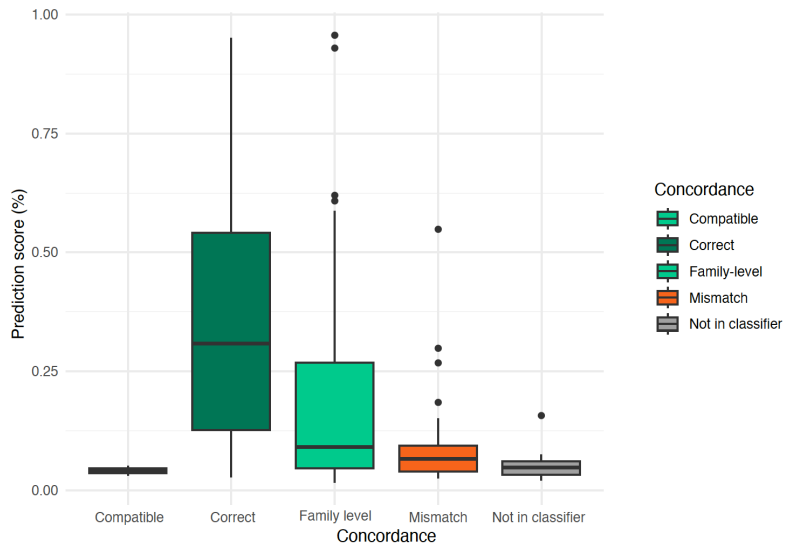


Figure 4-7 MNP-Flex scores for Rapid-CNS[2] data at different concordance levels (Reprinted from Patel et. al 2024[117])

Among all datasets tested, the Rapid-CNS[2] datasets showed the lowest confidence scores (**Figure 4-5**). Reassuringly, as indicated in **Figure 4-7**, correctly classified samples were assigned higher scores while incorrect predictions had low prediction scores. Thus, the model does not make incorrect predictions confidently. This is especially important for tools to be used in a medical context. I speculate that the reduced classification accuracy in these samples is largely attributable to the significant number of missing values present in the dataset. Missing data can disrupt the model's ability to make accurate predictions, particularly in complex classification tasks like ours involving methylation patterns. In comparing the Heidelberg and Nottingham datasets, I observed that the Heidelberg dataset had notably lower accuracy rates. This discrepancy appears to be linked to the higher frequency of missing values in the Heidelberg data, which can be traced to technical differences in sequencing approaches. Specifically, a greater proportion of samples in the Heidelberg cohort were run using R9 flow cells rather than the new R10 flow cells, which are known to yield higher accuracy data and lower error rates. Additionally, many of the Heidelberg samples were processed on GridION or MinION devices, which generally produce fewer reads and lower throughput compared to the PromethION device, used more frequently for Nottingham samples. These factors together likely contributed to the overall lower performance of the Heidelberg dataset.

I compared all samples run on the GridION in Heidelberg and Nottingham to further analyse the effect of flowcell configuration. I compared 95 samples sequenced with R9 flowcells to 116 samples sequenced with R10 flowcells.

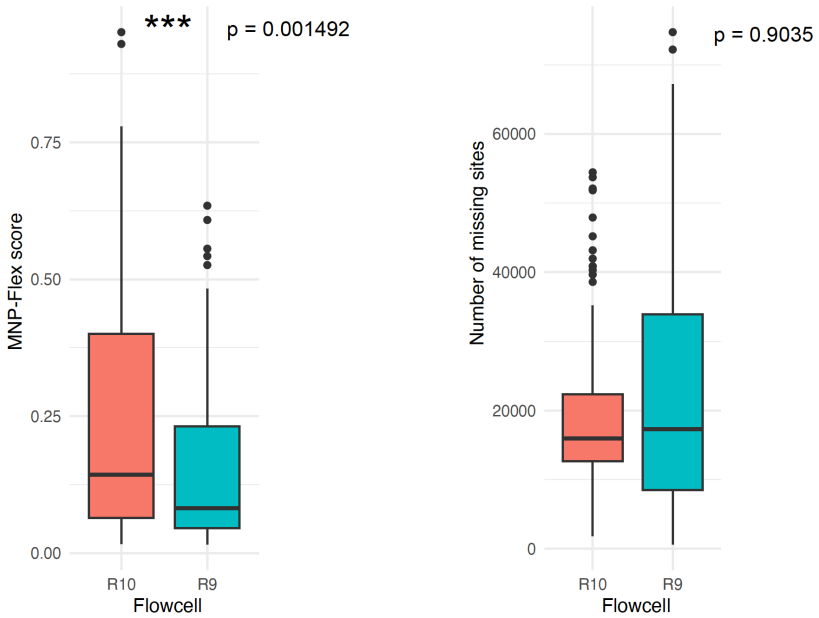


Figure 4-8 Comparison of R9 and R10 flowcells for samples sequenced on the GridION (Reprinted from Patel et. al 2024[117])

As shown in **Figure 4-8**, I found that there was a statistically significant difference (unpaired Wilcoxon test, p = 0.001492) in the MNP-Flex subclass prediction score with R10 flowcells tending to have higher scores. On the other hand, there was no significant difference in the number of missing sites (unpaired Wilcoxon test, p = 0.9035). Since the number of available sites remained similar, this indicated a significant increase in per CpG accuracy for R10 flowcells. Importantly, I observed that data generated exclusively using R10 flow cells demonstrated an improvement in classification accuracy for MNP-Flex. This finding underscores the impact of technical advancements, such as the transition from R9 to R10 flow cells, on the quality and reliability of methylation data, further supporting the notion that optimising sequencing technology can mitigate some of the issues associated with randomness in data and improve overall model performance.

4.3.5 Effect of tissue type

To estimate the effect of tissue type used, I compared the prediction scores across all samples. The dataset had 48,621 FFPE samples and 30,378 frozen samples sequenced using the different technologies. I performed a Wilcoxon rank sum test with continuity correction to compare the distribution of scores between frozen and FFPE samples. Additionally, I

employed the two one-sided test (TOST) procedure to assess equivalence within the predefined margin of [-0.01, 0.01][234].

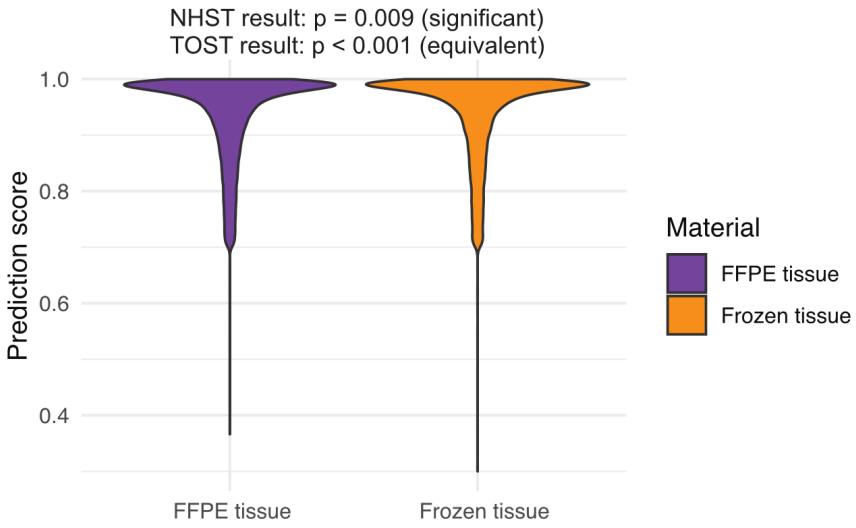


Figure 4-9 MNP-Flex prediction scores over tissue preparation types (Reprinted from Patel et. al 2024[117])

As shown in **Figure 4-9**, the results indicated a significant effect in both the equivalence test and the null hypothesis significance test (NHST). The equivalence test was significant (W = 654,773,271, p = 2.04e-118), indicating that the observed effect fell within the equivalence bounds. Similarly, the null hypothesis test was significant (W = 730,355,988, p = 0.009), leading me to reject the null hypothesis that the effect size is equal to zero. Further, the results from the TOST procedure supported these findings, with both the lower bound (W = 810,553,731, p < 0.001) and upper bound (W = 654,773,271, p < 0.001) tests showing strong statistical significance. This provided strong evidence that the observed effect is equivalent within the defined bounds. Despite these significant results, the effect size was notably small. The median of differences was -0.000891 with a 90% confidence interval of [-0.0014, -0.0003], suggesting that the absolute difference in model performance between the tissue types is minimal. Additionally, the rank-biserial correlation was -0.011034, with a 90% confidence interval of [-0.018, -0.0041], indicating a very weak relationship between the ranks of the frozen and FFPE data. Thus, we observed a significant effect (NHST), but this was significant due to the extremely high sample size of the several thousand samples and when tested for equivalence using TOST it was within the small margin range of 2 % score and thus the effect can be neglected.

This suggests that although the differences in scores between frozen and FFPE tissue are statistically significant, the practical impact of this difference is likely minimal.

4.4 Discussion

MNP-Flex expands the utility of methylation-based CNS tumour classification to include both array and sequencing-based data, maintaining high accuracy across multiple platforms. Its ability to handle missing data and deliver reliable predictions with a lower threshold makes it a versatile tool for clinical and research applications.

Methylation classification has evolved into a robust framework applicable to pan-CNS tumour classification. This is evidenced by the exponential increase in the number of classes, starting with the Sturm et. al defined glioma classes, progressing to the first CNS tumour classifier with 91 classes, and now expanding to encompass 184 classes in the latest version of the MNP classifier.[13,19,236]. The extension of this clinically relevant tool to platforms beyond methylation arrays opens up opportunities for broader application, enhances the classification scheme, and allows us to leverage the advantages of sequencing-based approaches to gain a deeper understanding of the differences between the classes and the potentially actionable mechanisms driving them.

The MNP-Flex classifier could successfully classify Rapid-CNS^2 samples providing granular classifications for the samples with 184 classes improving from the coarse classification provided by the in-built model with 91 classes. Cases that were unresolved due to absence of a reference class, could be successfully classified after MNP-Flex classification. This results in an integrated Rapid-CNS^2 workflow, as demonstrated in **Figure 4-10**, to report a coarse classification in the intraoperative time frame that is sufficient for resection decisions followed by the full spectrum of alterations including the granular methylation classification by MNP-Flex on the next day.

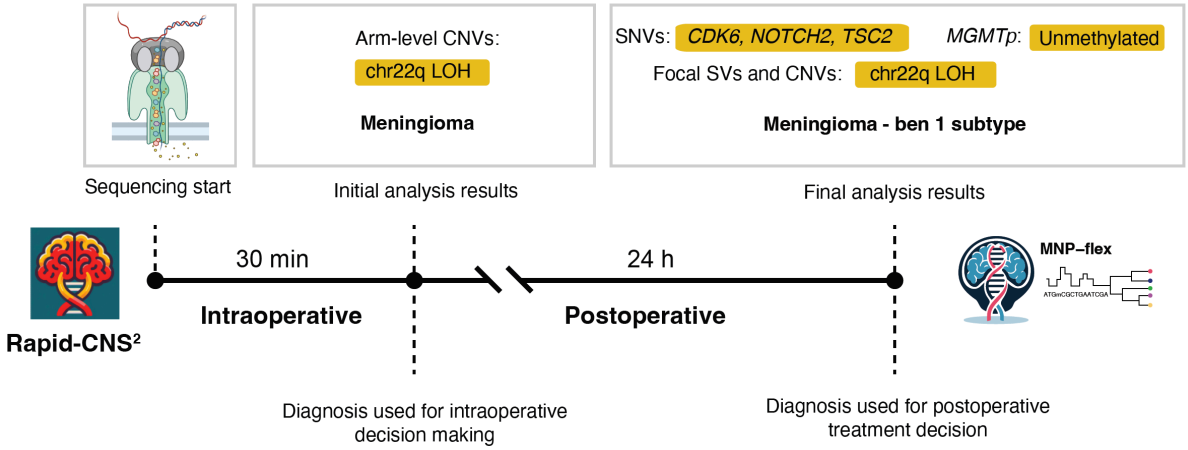


Figure 4-10 Example of the end-to-end workflow combining intraoperative and postoperative analysis (Reprinted from Patel et. al 2024[117])

MNP-Flex uses a gradient-boosted model and leverages the inherent ability of the architecture to handle missing values. However, as the results show, this poses limitations when dealing with sparse data that contains a high proportion of missing values. However, as the results indicate, dealing with sparse data containing a large proportion of missing values presents challenges. Consequently, we were unable to use MNP-Flex for classification intraoperatively in the Rapid-CNS[2] workflow, and could only apply it at the end of the sequencing run, once a substantial amount of data had been collected. This restricts the classifier's use to dense sequencing datasets such as high-coverage Nanopore WGS, WGBS, or methylation panels, making it unsuitable for classifying low-pass WGS or the error-prone Nanopore FFPE data. Although not described in this work, our attempts to train gradient-boosted models on data with varying degrees of missingness improved classification for sparse data at the cost of reduced scores and accuracy over dense data which deemed it unsuitable for a truly platform-agnostic approach. Errors in sequencing-based data arise either from random missing sites or from low coverage, which can result in incorrect methylation values. Improved classification may come from either refining the features or enhancing the model architecture and parameters[237]. Therefore, I speculate that a two-pronged approach—imputation of missing values followed by the use of a classifier capable of accounting for data errors—might be an effective strategy for this task. Given their ability to learn compact representations of data and reconstruct missing or noisy inputs, autoencoders offer a promising approach for imputing missing values in methylation data, where accurate reconstruction is crucial Autoencoders and their variants offer powerful deep learning tools for non-linear dimensionality reduction, clustering, data generation, imputation, and classification tasks[238-241]. Using autoencoders for

DNA methylation data could have the additional advantage of understanding how specific CpG sites relate through shared latent features, which could provide valuable insights into methylation signatures. Imputing such data before feeding it into MNP-Flex could already enhance classifier performance. In particular, masked autoencoders offer an elegant solution to handle missing or erroneous methylation values. A masked autoencoder randomly hides portions of the input data during training and attempts to reconstruct the masked information. This method not only encourages the model to learn robust representations of the underlying data but also simulates the real-world challenge of missing or erroneous CpG sites caused by low sequencing depth or random dropouts. Applying masked autoencoders with well-defined masking criteria could help simulate the inherent errors introduced by sequencing techniques, especially in low-quality datasets such as those derived from FFPE samples, which often exhibit high error rates in nanopore sequencing. Additionally, models could be trained on data with incrementally erroneous values to account for errors introduced by low coverage in sequencing-based data, as well as any inaccuracies from the imputation process. In practice, the model could be trained on partitions of the reference dataset with error rates of up to 15% to simulate real-world errors or higher to also account for the prohibitive error rates in nanopore sequencing of FFPE samples[242].To further mimic sequencing-based errors, incorrect sites could be positioned adjacent to one another, similar to erroneous methylation values caused by low-quality reads or reads originating from the normal compartment of the tumour. Such an approach could be combined with masked autoencoders, where predefined masking criteria could allow the model to learn from and correct for these localised sequencing errors. In this context, masked autoencoders would play a dual role: both as a tool for imputing missing values and as a mechanism for error correction in noisy methylation data. This approach not only improves the classifier's performance by providing cleaner input data but also enables a deeper understanding of the underlying methylation landscape by identifying patterns that emerge from noisy or incomplete data. Thus, the integration of masked autoencoders could be a valuable addition to the workflow for methylation-based classification, improving both the robustness and accuracy of the predictions in a real-world clinical setting.

Another promising area of application for MNP-Flex is liquid biopsies. Detecting cancer types using non-invasive methods, such as liquid biopsies from blood or cerebrospinal fluid (CSF), could be a game-changer for CNS tumours, particularly in cases where surgery is either impossible or too risky. Circulating tumour DNA (ctDNA) is found in the blood of fewer than 10% of glioma patients, as the blood-brain barrier restricts the release of biomarkers like cell-

free DNA (cfDNA) from brain tumours into the bloodstream, making detection with conventional assays challenging[243-246]. The shedding of cfDNA into the CSF is increased in patients with high tumour burden, progressive or metastasised tumours, or tumours adjacent to ventricles. Liquid biopsies provide a valuable opportunity to reassess the tumour molecular profile in cases of recurrence and track its evolution during treatment by detecting mutation fractions in plasma and CSF. Additionally, longitudinal plasma-based liquid biopsies may complement neuroimaging in evaluating treatment response, particularly in distinguishing between pseudoprogression and true progression—an important clinical issue highlighted by the RANO (Response Assessment in Neuro-Oncology) consortium's liquid biopsy task force[247]. A study by Nassiri et. *al* demonstrated successful detection and classification of methylomes of cfDNA from plasma for intracranial tumours using a immunoprecipitation approach[245]. Previous studies have reported detection of pathogenic mutations, CNVs methylation classification using NGS DNA sequencing as well as nanopore sequencing albeit for restricted set of CNS tumour types[248-251]. A major challenge with liquid biopsy data is to distinguish tumour DNA from healthy DNA. This could potentially be solved by borrowing from the field of Natural Language Processing (NLP) with models like BERT (Bidirectional Encoder Representations from Transformers). A recent study has proposed MethylBERT, a BERT based method for read-level methylation pattern identification and tumour purity estimation[252]. Integrating such a model with MNP-Flex would enable processing of read-level methylomes and classifying sequencing reads as either tumour or normal based on their methylation patterns. This prior sorting of tumour reads would enhance the precision of MNP-Flex for cell-free DNA from plasma or CSF. A considerable proportion of errors by MNP-Flex in this work were mispredictions as tumour microenvironment or reactive tissue resulting from low tumour purity. The MethylBERT approach could also be useful for improving classification accuracy in such cases.

The accumulation of sequencing-based methylation profiles alongside methylation classification using MNP-Flex offers the additional advantage of exploring genetic data in conjunction with epigenetic information. This integrative approach holds significant potential for uncovering subtle distinctions among various tumour subtypes, especially novel subtypes. Particularly with Nanopore sequencing, the simultaneous availability of mutational and methylation data from the same read provides a unique opportunity to investigate the interplay between genetic mutations and epigenetic modifications. This could be especially valuable in understanding the complex drivers of CNS tumours, where both genetic and epigenetic factors

contribute to tumour behaviour and progression. By combining these data, the community can gain deeper insights into tumour heterogeneity, enabling more precise diagnostics and potentially identifying novel therapeutic targets. This integrated analysis could further enhance our understanding of tumour evolution, treatment resistance, and the discovery of new molecular markers, ultimately paving the way for more personalised approaches in cancer management.

Another limitation of MNP-Flex and methylation classifiers, in general, is the lack of representation from ethnically diverse populations in the reference dataset, which is predominantly derived from patients in the Global North. It is well-known that different diseases can manifest differently across ethnic groups, with variations in genetic and epigenetic drivers contributing to these differences[253,254]. The role of ethnicity in disease manifestation is particularly well-documented in fields like cancer, cardiovascular diseases, and autoimmune disorders. Studies have highlighted how genetic variants, allele frequencies, and environmental interactions vary across populations, influencing disease susceptibility, progression, and response to treatment[255-257]. Thus, it follows that the classes that form the basis of MNP-Flex might not accurately reflect the tumour subtypes in the underrepresented populations. This underscores the need for an ethnically inclusive approach to ensure healthcare interventions are tailored to diverse populations. Incorporating data from a broader range of ethnic groups would improve MNP-Flex to make it not only platform-agnostic but truly a flexible classifier catering to diverse technologies as well as individuals.

Chapter 5 Discussion and outlook

The WHO CNS5 classification reflects a significant shift towards a more precise, molecularly driven approach to CNS tumour diagnosis. Molecular testing is now essential for defining many tumour types, providing critical diagnostic and prognostic information that histology alone cannot offer. From *IDH* mutations in gliomas, *CDKN2A/B* deletions in meningiomas to gene fusions in ependymomas and methylation profiles in medulloblastomas, molecular markers have become the cornerstone of CNS tumour classification. This move towards molecular diagnostics has enhanced diagnostic accuracy and also opened the door for personalised treatments, leading to improved patient outcomes in neuro-oncology.

Traditional molecular diagnostics in neuropathology have long been hindered by cumbersome methods and significant financial barriers, preventing several patients from accessing timely and precise diagnoses. The centralisation of testing, while rational considering the cost and effort of running such facilities, has further contributed to delays, straining resources and extending turnaround times. Building on the momentum of the WHO CNS5 classification, the development of tools like Rapid-CNS^2 and MNP-Flex offers a response to these challenges in face of the growing need for molecularly driven diagnostics in CNS tumours. As molecular markers become central to diagnostic and treatment decisions, the ability to rapidly and accurately assess these markers is critical. Rapid-CNS^2, with its adaptive sampling and nanopore sequencing technology, enables comprehensive molecular profiling, including methylation classification, within clinically relevant and even intraoperative timeframes. MNP-Flex further complements this by providing platform-agnostic methylation classification compatible with the latest versions of the MNP classifier, ensuring that updated diagnostic classes can be identified regardless of the sequencing method used. This work presents a complete end-to-end streamlined workflow as demonstrated in **Figure 5-1**, with the frozen section being sequenced intraoperatively to report a broad v11 methylation classification and copy number profile in a surgical timeframe, followed by next day reporting of the full spectrum

of molecular alterations along with granular v12 based methylation classification using MNP-Flex.

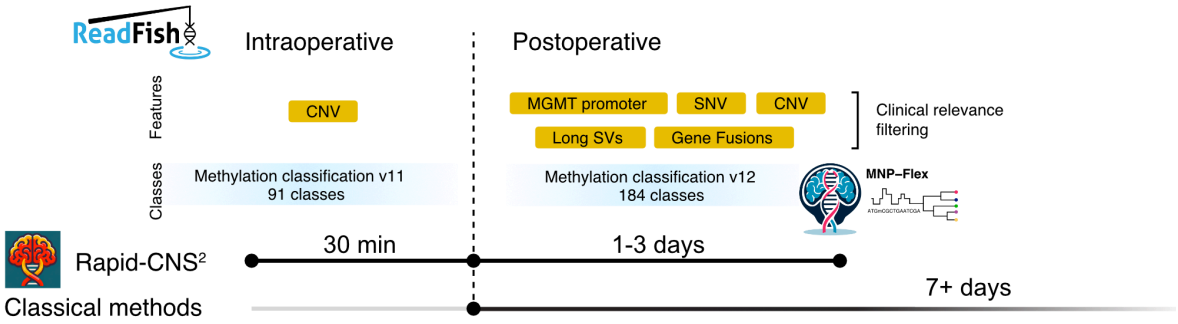


Figure 5-1 Overview of the end-to-end pipeline

The tools developed in this work are catered to support the advances in CNS tumour diagnostics and the ongoing shift towards precision medicine in neuro-oncology. Moreover, the tools are readily adaptable, making them well-suited for the constantly increasing number of molecular criteria and evolving technologies. For instance, BGI has unveiled CycloneSeq, its own nanopore-based sequencing approach[258]. Methylation data generated using this approach could easily be used with MNP-Flex without retraining. Additionally, newly discovered methylation subytpes can be easily incorporated to retrain MNP-Flex, while new molecular targets could be added to the panel for Rapid-CNS[2]. Importantly, these tools are not limited to CNS tumours; they can be adapted to various tumour types making them a flexible foundation for broader cancer diagnostics. Other methylation classifiers like sarcoma, sinonasal and any classifiers that might be developed in the future could also be trained using the MNP-Flex model in combination with targeting using adaptive sampling[259,260].

One of the most practical consequences of Rapid-CNS[2] is the dramatically reduced turnaround time. Comprehensive molecular diagnostics, including targeted regions, can now be completed in less than two days, and methylation classification and broad CNV profiling can be performed in under 30 minutes. This is a substantial improvement from the average 20-day wait time associated with conventional methods (**Figure 5-1**). From a postoperative management perspective, this accelerates the timeline to enable faster treatment decisions, allowing for timely access to molecularly informed disease management, in distant or resource-limited facilities.

A recent report in from the UK found that less than 5% patients with a brain tumour have access to WGS, even though all individuals with brain tumours are eligible for WGS[261]. The report by the Tessa Jowell Brain Cancer Mission specifically stated the Rapid-CNS^2 based approach adopted at University of Nottingham as a future technology that should be considered to improve access to precision diagnostics for all patients. Although nanopore devices are not yet widely available in all neurosurgical settings, the affordability of the technology—where the smallest device capable of running Rapid-CNS^2 costs about one-fiftieth of the setup required for conventional methylation testing—positions it for swift adoption.

Rapid-CNS^2 leverages single-molecule sequencing which enables detection of both genetic and epigenetic modifications of the same molecule. By integrating data on mutations, CNVs, and methylation—especially through long-read sequencing—this approach holds significant promise for accurately identifying subclonal reads[119,262,263]. A robust MNP-Flex model could further extend this capability to identify subclonal methylation classes, providing critical insights into tumour heterogeneity for potential targeted therapies. As nanopore sequencing technology continues to advance, along with improvements in clonotyping and methylation classification models, the future could see routine non-invasive liquid biopsy for disease monitoring[264,265]. Such approaches would enable the detection of clonal diversity, the targeting of emerging clones, and the ability to assess treatment efficacy by accurately estimating tumour burden[266,267].

The successful use of MNP-Flex across global datasets underscores its generalisability and utility in diverse diagnostic and research settings. The static nature of MNP-flex, as opposed to the dynamic ad-hoc classification of Rapid-CNS^2, allows it to cover the full granularity of the MNP v12 classifier with an even lower computational footprint. This has important implications not only for future iterations of methylation classifiers but also for regulatory compliance, as dynamic systems often face more scrutiny in clinical applications. With improvements, MNP-Flex could replace the ad-hoc model in intraoperative settings as well, providing an even more streamlined diagnostic process.

A further step to accessibility would be to replace molecular testing with solely H&E based analysis in certain cases. The last decade has seen an exponential rise in the development of neural network and transformer-based AI models to predict the occurrence of molecular markers like mutations, gene expression, molecular signatures and other prognostically relevant features directly from a H&E slide[246,268-276]. A recent study from NCI has demonstrated the prediction of about 10 broad CNS tumour methylation subtypes from H&E subtypes[277]. A

major limitation to training such models is the lack of sufficient annotated data to fully exploit the deep networks that are dominating other computer vision fields. To overcome this, a number of foundation models have been released that can serve as pre-training followed by fine tuning on specific data and labels[271,273,274,278]. Harnessing such foundation models and the large annotated dataset from the Department of Neuropathology, University Hospital Heidelberg and co-supervision by Felix Sahm and Moritz Gerstung, I am currently working on developing and evaluating a transformer-based model that can predict 108 methylation-based subtypes of CNS tumours. As molecular classification guidelines evolve, such tools will be essential for democratising access to high-quality, cutting-edge diagnostics.

The development of Rapid-CNS^2 and MNP-flex took place in a fast-moving technological environment, where sequencing devices, analytical tools, and methodologies were constantly evolving. Throughout this process, the tools were continuously refined to incorporate the latest advancements, such as improved adaptive sampling with updates to ReadFish and modifications in nanopore devices. This adaptability ensured that the tools stayed up-to-date with current methods, offering an example of how to build cutting-edge systems capable of keeping pace with technological advancements. Beyond their immediate diagnostic utility, Rapid-CNS^2 and MNP-Flex provide a structure for future developments in molecular diagnostics. Their modular design allows them to evolve in tandem with emerging technologies and discoveries, positioning them as platforms that can be expanded to accommodate the complexities of different tumour types. As more tumour-specific data becomes available, these tools can be further adapted. Moreover, the framework they provide serves as a model for the development of new diagnostic tools that can quickly integrate advancements in sequencing technology and computational analysis.

While tools like Rapid-CNS^2 and MNP-Flex demonstrate immense potential for both research and clinical applications, it is important to recognise that these tools are currently for research use only (RUO). Despite their proven efficacy in molecular diagnostics, their dedicated implementation in clinical practice hinges on obtaining the necessary clinical certifications. One such certification is the In Vitro Diagnostic Regulation (IVD-R) in Europe, which ensures that medical devices, including diagnostic tools, meet stringent safety and performance standards before being deployed in healthcare settings. Achieving IVD-R certification, or similar regulatory approvals in other regions, is essential for the formal integration of new workflows into routine clinical workflows. This process would involve rigorous validation studies, demonstrating reproducibility, reliability, and accuracy across diverse patient

populations and clinical environments. Securing clinical certification would unlock the potential of these tools for use in diagnosing CNS tumours and other cancers in a real-world clinical setting, allowing for more widespread adoption. The transition from RUO to certified clinical diagnostics is a critical next step in bringing the benefits of advanced molecular profiling and precision medicine to patients. As these tools continue to evolve, the pursuit of clinical certification will ensure they can contribute meaningfully to improving patient outcomes on a global scale. Efforts are already underway in Heidelberg to bridge the gap between research and clinical application. Heidelberg Epignostix GmbH, a DKFZ and UKHD spin-off, is focused on obtaining IVD-R certification and securing market access for the MNP methylation classifier while also developing further tools for clinical use[275]. This is a crucial step to ensuring that such tools can be used in clinical settings, where their potential can be fully realised.

Current tools for methylation classification of CNS tumours like the MNP classifier and similarly Rapid-CNS[2] and MNP-Flex, suffer from a severe lack of diversity, as 90% of genomic data used to develop these tools comes from individuals of European and North American descent who form 17% of the global population[278]. This overlooks significant global populations, particularly those from low-to-middle-income countries (LMICs) in Asia, Africa, and South America. As a result, the biological insights derived from these tools are incomplete, failing to capture the unique genetic and epigenetic variations present in underrepresented populations. This gap not only limits our understanding of disease in these regions but also hampers the global applicability of the classifiers. The MNP Outreach Consortium, a Hopp Children's Cancer Centre (KiTZ) and UKHD initiative, represents a transformative approach to addressing this disparity by extending access to methylation profiling and diagnostic technologies to LMICs[279]. By partnering with multiple centres across the global south, the consortium aims to close the diversity gap in genomic databases, fostering more inclusive and representative research. This initiative has the potential to uncover novel genetic markers that are prevalent in LMICs, contributing to a more equitable global health landscape.

In alignment with my broader goal of making molecular diagnostics more accessible, my next step will be to work for Heidelberg Epignostix to develop approaches for new technologies and support the MNP Outreach consortium in implementing them in centres located in the global south. Through this role, I will be directly involved in translating these advanced technologies from research tools into clinically certified diagnostics, ensuring that they can be implemented in real-world healthcare settings. This endeavour is not just a professional advancement but part of my ongoing mission to reduce the barriers to molecular diagnostics. By also focusing

on certification and market access, I hope to make precision diagnostics more widely available, particularly in under-resourced regions.

Publications and presentations

During the course of my doctoral thesis, I (co-) authored the following publications, patents and presented at the following conferences or seminars:

Publications

Areeba Patel, Kirsten Göbel, …. Martin Sill, Felix Sahm. 2024-04-10. "Versatile, Accessible Cross-Platform Molecular Profiling of Central Nervous System Tumors: Web-Based, Prospective Multi-Center Validation." *Research Square* (as of 2024-09-29, under revision at *Nature Medicine*) https://dx.doi.org/10.21203/rs.3.rs-4182910/v1.

Areeba Patel*, Helin Dogan*, Alexander Payne, Elena Krause, and others. 2022-03-31. "Rapid-CNS2: Rapid Comprehensive Adaptive Nanopore-Sequencing of CNS Tumors, a Proof-of-Concept Study." *Acta Neuropathologica*, 143(5). https://dx.doi.org/10.1007/s00401-022-02415-6.

Deacon, S., ... **A. Patel**, R. Goldspring, S. Brandner, F. Sahm, S. Smith, SML Paine, and M. Loose. 2024-09-11. "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." *medRxiv*. https://dx.doi.org/10.1101/2024.09.10.24313398.

Skarphedinn Halldorsson, ... **Areeba Patel**, Petter Brandal, and others. 2024-06-01. "Neuropathology and Applied Neurobiology | BNS Journal | Wiley Online Library." *Neuropathology and Applied Neurobiology*, 50(3). https://dx.doi.org/10.1111/nan.12984.

Iser, Florian, ... **Areeba Patel**, Duy Nguyen, Leon D. Kaulen, and others. 2024-07-15. "Cerebrospinal Fluid cfDNA Sequencing for Classification of Central Nervous System Glioma." *Clinical Cancer Research*, 30(14). https://dx.doi.org/10.1158/1078-0432.CCR-23-2907.

Helin Dogan, Christina Blume, **Areeba Patel**, Gerhard Jungwirth, Lisa Sogerer, and others. 2022-08-19. "Single-Cell DNA Sequencing Reveals Order of Mutational Acquisition in TRAF7/AKT1 and TRAF7/KLF4 Mutant Meningiomas." *Acta Neuropathologica*, 144(4). https://dx.doi.org/10.1007/s00401-022-02485-6.

Miriam Ratliff, ... **Areeba Patel**, Elena Maier, Loic Cousin, and others. 2022-06-12. "Patient-Derived Tumor Organoids for Guidance of Personalized Drug Therapies in Recurrent Glioblastoma." *International Journal of Molecular Sciences*, 23(12), 6572. https://dx.doi.org/10.3390/ijms23126572.

Christina Blume, ... **Areeba Patel**, Matthias Schlesner, and Felix Sahm. 2021-05-11. "Integrated Phospho-Proteogenomic and Single-Cell Transcriptomic Analysis of Meningiomas Establishes Robust Subtyping and Reveals Subtype-Specific Immune Invasion." *bioRxiv*. https://dx.doi.org/10.1101/2021.05.11.443369.

Patent

DKFZ, University Hospital Heidelberg, University of Nottingham. Rapid comprehensive adaptive nanopore-sequencing of CNS tumors, a proof of concept study (RAPID-CNS2) (filings EP21190233.3, PCT/EP2022/072034 published under EP4131274)

Oral Presentations

Invited speaker at the *Society for Neurooncology Sub-Saharan Africa (SNOSSA) First Thursday* webinar, September 2024

Invited speaker at the *UK Genome Science meeting*, Bristol (United Kingdom), July 2024

Plenary presentation at *International Society of Paediatric Neurooncology (ISPNO) Annual Meeting*, Philadelphia (United States of America), June 2024

Invited speaker at the *British Neuropathological Society Summer School,* Cirencester (United Kingdom), June 2024

Oral presentation at the *Society of Neurooncology (SNO) Annual Meeting*, Vancouver (Canada), November 2023

Plenary presentation and Best Oral Presentation Award for Translational research at the *Annual Meeting of the European Association of Neuro-oncology (EANO)*, Rotterdam (Netherlands), September 2023

Oral presentation at *Neurowoche*, Berlin (Germany), November 2022

Plenary presentation at *Nanopore Community Meeting*, Virtual, November 2021

Poster Presentations

Poster presentation at the *Association of Molecular Pathology (Europe)* conference, Madrid (Spain), June 2024

Poster presentation at *London Calling*, London (United Kingdom), May 2024

Poster presentation at the *Society of Neurooncology (SNO) Annual Meeting*, Tampa (United States of America), November 2022

Poster presentation at *Society of Neurooncology (SNO) Annual Meeting*, Virtual, November 2021

References

1. RE, F. & RG, G. Molecular configuration in sodium thymonucleate - PubMed. *Nature* **171**(04/25/1953).
2. JD, W. & FH, C. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid - PubMed. *Nature* **171**(04/25/1953).
3. AM, M. & W, G. A new method for sequencing DNA - PubMed. *Proceedings of the National Academy of Sciences of the United States of America* **74**(1977 Feb).
4. Sanger, F., Nicklen, S. & Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**(1977/12).
5. LM, S.*, et al.* Fluorescence detection in automated DNA sequence analysis - PubMed. *Nature* **321**(1986 Jun).
6. JD, W. & RM, C.-D. Origins of the Human Genome Project - PubMed. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **5**(1991 Jan).
7. Birney, E. The International Human Genome Project. *Human Molecular Genetics* **30**(2021/10/10).
8. M, R., M, U. & P, N. A sequencing method based on real-time pyrophosphate - PubMed. *Science (New York, N.Y.)* **281**(07/17/1998).
9. P, N. Enzymatic method for continuous monitoring of DNA polymerase activity - PubMed. *Analytical biochemistry* **167**(1987 Dec).
10. Gaynor, S.M.*, et al.* Yield of genetic association signals from genomes, exomes and imputation in the UK Biobank. *Nature Genetics 2024* (2024-09-25).
11. Network, T.C.G.A.R. Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *https://doi.org/10.1056/NEJMoa1402121* **372**(2015).
12. Sj, C., J, H., Cl, P. & M, F. High sensitivity mapping of methylated cytosines - PubMed. *Nucleic acids research* **22**.
13. Sturm, D.*, et al.* Hotspot Mutations in H3F3A and IDH1 Define Distinct Epigenetic and Biological Subgroups of Glioblastoma. *Cancer Cell* **22**(2012/10/16).
14. Hovestadt, V.*, et al.* Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. *Nature 2014 510:7506* **510**(2014-05-18).
15. PD, J.*, et al.* Atypical Teratoid/Rhabdoid Tumors Are Comprised of Three Epigenetic Subgroups with Distinct Enhancer Landscapes - PubMed. *Cancer cell* **29**(03/14/2016).
16. JT, B.*, et al.* DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines - PubMed. *Genome biology* **12**(2011).
17. Sandoval, J.*, et al.* Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* **6**(2011-6-1).
18. Pidsley, R.*, et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biology 2016 17:1* **17**(2016-10-07).

19. Capper, D.*, et al.* DNA methylation-based classification of central nervous system tumours. *Nature 2018 555:7697* **555**(2018-03-14).

20. Noguera-Castells, A., García-Prieto, C.A., Álvarez-Errico, D. & Esteller, M. Validation of the new EPIC DNA methylation microarray (900K EPIC v2) for high-throughput profiling of the human DNA methylome. *Epigenetics* **18**(2023-12-31).

21. Lizio, M.*, et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biology 2015 16:1* **16**(2015-01-05).

22. An integrated encyclopedia of DNA elements in the human genome - PubMed. *Nature* **489**(09/06/2012).

23. Kasianowicz, J.J.*, et al.* Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences* **93**(1996).

24. Deamer, D.*, et al.* Three decades of nanopore sequencing. *Nature Biotechnology 2016 34:5* **34**(2016).

25. Branton, D.*, et al.* The potential and challenges of nanopore sequencing. *Nature Biotechnology 2008 26:10* **26**(2008).

26. Wang, Y.*, et al.* Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology 2021 39:11* **39**(2021-11-08).

27. Wang, Y.*, et al.* Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology 2021 39:11* **39**(2021).

28. Lv, X., Chen, Z., Lu, Y. & Yang, Y. An End-to-end Oxford Nanopore Basecaller Using Convolution-augmented Transformer. *bioRxiv* (2020-11-10).

29. Zeng, J.*, et al.* Frontiers | Causalcall: Nanopore Basecalling Using a Temporal Convolutional Network. *Frontiers in Genetics* **10**(2020).

30. Yuen, Z.W.-S.*, et al.* Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing. *Nature Communications 2021 12:1* **12**(2021).

31. Rang, F.J.*, et al.* From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biology 2018 19:1* **19**(2018).

32. Pagès-Gallego, M., de Ridder, J., Pagès-Gallego, M. & de Ridder, J. Comprehensive benchmark and architectural analysis of deep learning models for nanopore sequencing basecalling. *Genome Biology 2023 24:1* **24**(2023).

33. Lv, X., Chen, Z., Lu, Y. & Yang, Y. An End-to-end Oxford Nanopore Basecaller Using Convolution-augmented Transformer. *bioRxiv* (2020).

34. Konishi, H., Yamaguchi, R., Yamaguchi, K., Furukawa, Y. & Imoto, S. Halcyon: an accurate basecaller exploiting an encoder–decoder model with monotonic attention. *Bioinformatics* **37**(2021).

35. Huang, N., Nie, F., Ni, P., Luo, F. & Wang, J. SACall: A Neural Network Basecaller for Oxford Nanopore Sequencing Data Based on Self-Attention Mechanism | IEEE Journals & Magazine | IEEE Xplore. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **19**(2022).

36. Nurk, S.*, et al.* The complete sequence of a human genome. *bioRxiv* (2021).

37. Jain, M.*, et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology 2018 36:4* **36**(2018).

38. Pugh, J. The Current State of Nanopore Sequencing. *Methods in Molecular Biology* (2023).

39. Stevanovski, I.*, et al.* Comprehensive genetic diagnosis of tandem repeat expansion disorders with programmable targeted nanopore sequencing. *Science Advances* **8**(2022).

40. P, G.*, et al.* Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing - PubMed. *Nature biotechnology* **37**(2019).

41. A Nanopore Sequencing-Based Assay for Rapid Detection of Gene Fusions - PubMed. *The Journal of molecular diagnostics : JMD* **21**(2019).

42. Schmidt, T.T.*, et al.* High resolution long-read telomere sequencing reveals dynamic mechanisms in aging and cancer. *Nature Communications 2024 15:1* **15**(2024).

43. Xu, L., Seki, M., Xu, L. & Seki, M. Recent advances in the detection of base modifications using the Nanopore sequencer. *Journal of Human Genetics 2019 65:1* **65**(2019).

44. Leger, A.*, et al.* RNA modifications detection by comparative Nanopore direct RNA sequencing. *Nature Communications 2021 12:1* **12**(2021).

45. M, P.*, et al.* Nanopore sequencing of single-cell transcriptomes with scCOLOR-seq - PubMed. *Nature biotechnology* **39**(2021 Dec).

46. K, L., V, M., P, B. & R, W. High throughput error corrected Nanopore single cell transcriptome sequencing - PubMed. *Nature communications* **11**(08/12/2020).

47. Zhang, J.-Y.*, et al.* A single-molecule nanopore sequencing platform. *bioRxiv* (2024).

48. Ying, Y.-L.*, et al.* Nanopore-based technologies beyond DNA sequencing. *Nature Nanotechnology 2022 17:11* **17**(2022).

49. Motone, K.*, et al.* Multi-pass, single-molecule nanopore reading of long protein strands. *Nature 2024* (2024).

50. Ferlay, J.*, et al.* Global Cancer Observatory: Cancer Today. (2020).

51. D, H. Hallmarks of Cancer: New Dimensions - PubMed. *Cancer discovery* **12**(2022 Jan).

52. Hanahan, D. & Weinberg, R.A. The Hallmarks of Cancer. *Cell* **100**(2000).

53. D, H. Hallmarks of Cancer: New Dimensions - PubMed. *Cancer discovery* **12**(2022).

54. C, C., A, F. & V, R. Advances in the molecular classification of pediatric brain tumors: a guide to the galaxy - PubMed. *The Journal of pathology* **251**(2020 Jul).

55. Patil, P.A. & Giridhar, P. Epidemiology and Demography of Brain Tumors. *Evidence based practice in Neuro-oncology* (2021).

56. Louis, D.N.*, et al.* The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro-Oncology* **23**(2021/08).

57. Sahm, F.*, et al.* Molecular diagnostic tools for the World Health Organization (WHO) 2021 classification of gliomas, glioneuronal and neuronal tumors; an EANO guideline. *Neuro-Oncology* **25**(2023/10/03).

58. Bale, T.A. & Rosenblum, M.K. The 2021 WHO Classification of Tumors of the Central Nervous System: An update on pediatric low-grade gliomas and glioneuronal tumors. *Brain Pathology* **32**(2022/07/01).

59. Park, Y.W.*, et al.* The 2021 WHO Classification for Gliomas and Implications on Imaging Diagnosis: Part 1—Key Points of the Fifth Edition and Summary of Imaging Findings on Adult-Type Diffuse Gliomas. *Journal of Magnetic Resonance Imaging* **58**(2023/09/01).

60. Pfister, S.M.*, et al.* A Summary of the Inaugural WHO Classification of Pediatric Tumors: Transitioning from the Optical into the Molecular Era. *Cancer Discovery* **12**(2022/02/02).

61. Capper, D.*, et al.* EANO guideline on rational molecular testing of gliomas, glioneuronal, and neuronal tumors in adults for targeted therapy selection. *Neuro-Oncology* **25**(2023/05/04).

62. Capper, D.*, et al.* DNA methylation-based classification of central nervous system tumours. *Nature 2018 555:7697* **555**(2018).

63. Sturm, D.*, et al.* Multiomic neuropathology improves diagnostic accuracy in pediatric neuro-oncology. *Nature Medicine 2023 29:4* **29**(2023).

64. Capper, D.*, et al.* Practical implementation of DNA methylation and copy-number-based CNS tumor diagnostics: the Heidelberg experience. *Acta Neuropathologica 2018 136:2* **136**(2018).

65. Jaunmuktane, Z.*, et al.* Methylation array profiling of adult brain tumours: diagnostic outcomes in a large, single centre. *Acta Neuropathologica Communications 2019 7:1* **7**(2019).

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

67. Pickles, J.C.*, et al.* DNA methylation-based profiling for paediatric CNS tumour diagnosis and treatment: a population-based study. *The Lancet Child & Adolescent Health* **4**(2020/02/01).

68. Priesterbach-Ackley, L.P.*, et al.* Brain tumour diagnostics using a DNA methylation-based classifier as a diagnostic support tool. *Neuropathology and Applied Neurobiology* **46**(2020).

69. White, C.L.*, et al.* Implementation of DNA Methylation Array Profiling in Pediatric Central Nervous System Tumors: The AIM BRAIN Project: An Australian and New Zealand Children's Haematology/Oncology Group Study. *The Journal of Molecular Diagnostics* **25**(2023/10/01).

70. Weller, M.*, et al.* EANO guidelines on the diagnosis and treatment of diffuse gliomas of adulthood. *Nature Reviews Clinical Oncology 2020 18:3* **18**(2020).

71. Gonzalez Castro, L.N. & Wesseling, P. The cIMPACT-NOW updates and their significance to current neuro-oncology practice. *Neuro-Oncology Practice* **8**(2021).

72. Sahm, F.*, et al.* Molecular diagnostic tools for the World Health Organization (WHO) 2021 classification of gliomas, glioneuronal and neuronal tumors; an EANO guideline. *Neuro-Oncology* **25**(2023).

73. Louis, D.N.*, et al.* The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro-Oncology* **23**(2021).

74. Bale, T.A. & Rosenblum, M.K. The 2021 WHO Classification of Tumors of the Central Nervous System: An update on pediatric low-grade gliomas and glioneuronal tumors. *Brain Pathology* **32**(2022).

75. Bertero, L.*, et al.* Molecular neuropathology: an essential and evolving toolbox for the diagnosis and clinical management of central nervous system tumors. *Virchows Archiv 2023* (2023).

76. Maas, S.L.N.*, et al.* Integrated Molecular-Morphologic Meningioma Classification: A Multicenter Retrospective Analysis, Retrospectively and Prospectively Validated. *Journal of Clinical Oncology* **39**(2021).

77. F, S.*, et al.* DNA methylation-based classification and grading system for meningioma: a multicentre, retrospective analysis - PubMed. *The Lancet. Oncology* **18**(2017 May).

78. Choudhury, A.*, et al.* Meningioma DNA methylation groups identify biological drivers and therapeutic vulnerabilities. *Nature Genetics 2022 54:5* **54**(2022-05-09).

79. Hovestadt, V.*, et al.* Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. *Nature 2014 510:7506* **510**(2014).

80. C, N.*, et al.* An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma - PubMed. *Cell* **178**(08/08/2019).

81. E, B.-C.*, et al.* Tumor heterogeneity and tumor-microglia interactions in primary and recurrent IDH1-mutant gliomas - PubMed. *Cell reports. Medicine* **4**(11/21/2023).

82. Turcan, S.*, et al.* IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature 2012 483:7390* **483**(2012-02-15).

83. Venneti, S.*, et al.* Histone 3 Lysine 9 Trimethylation Is Differentially Associated With Isocitrate Dehydrogenase Mutations in Oligodendrogliomas and High-Grade Astrocytomas. *Journal of Neuropathology & Experimental Neurology* **72**(2013/04/01).

84. Baylin, S.B. & Jones, P.A. Epigenetic Determinants of Cancer. *Cold Spring Harbor Perspectives in Biology* **8**(2016-09-01).

85. Sturm, D.*, et al.* Hotspot Mutations in H3F3A and IDH1 Define Distinct Epigenetic and Biological Subgroups of Glioblastoma. *Cancer Cell* **22**(2012).

86. Sturm, D.*, et al.* Paediatric and adult glioblastoma: multiform (epi)genomic culprits emerge. *Nature Reviews Cancer 2014 14:2* **14**(2014).

87. Pd, J.*, et al.* Atypical Teratoid/Rhabdoid Tumors Are Comprised of Three Epigenetic Subgroups with Distinct Enhancer Landscapes - PubMed. *Cancer cell* **29**.

88. Guo, M.*, et al.* Epigenetic heterogeneity in cancer. *Biomarker Research 2019 7:1* **7**(2019).

89. Jaffe, A.E., Irizarry, R.A., Jaffe, A.E. & Irizarry, R.A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biology 2014 15:2* **15**(2014-02-04).

90. Esteller, M. Epigenetics in Cancer. *New England Journal of Medicine* **358**(2008-03-13).

91. Sahm, F.*, et al.* Next-generation sequencing in routine brain tumor diagnostics enables an integrated diagnosis and identifies actionable targets. *Acta Neuropathologica 2015 131:6* **131**(2015).

92. Stichel, D.*, et al.* Routine RNA sequencing of formalin-fixed paraffin-embedded specimens in neuropathology diagnostics identifies diagnostically and therapeutically relevant gene fusions. *Acta Neuropathologica 2019 138:5* **138**(2019).

93. Grassl, N.*, et al.* A H3K27M-targeted vaccine in adults with diffuse midline glioma. *Nature Medicine 2023 29:10* **29**(2023).

94. T, S.*, et al.* A vaccine targeting mutant IDH1 induces antitumour immunity - PubMed. *Nature* **512**(08/21/2014).

95. N2M2 (NOA-20) phase I/II trial of molecularly matched targeted therapies plus radiotherapy in patients with newly diagnosed non-MGMT hypermethylated glioblastoma - PubMed. *Neuro-oncology* **21**.

96. Watson, S.S.*, et al.* Fibrotic response to anti-CSF-1R therapy potentiates glioblastoma recurrence. *Cancer Cell* **42**(2024).

97. Bertero, L.*, et al.* Molecular neuropathology: an essential and evolving toolbox for the diagnosis and clinical management of central nervous system tumors. *Virchows Archiv 2023 484:2* **484**(2023-09-02).

98. M, K.*, et al.* Diffuse high-grade gliomas with H3 K27M mutations carry a dismal prognosis independent of tumor location - PubMed. *Neuro-oncology* **20**(01/10/2018).

99. EM, T.*, et al.* Prognostic value of medulloblastoma extent of resection after accounting for molecular subgroup: a retrospective integrated clinical and molecular analysis - PubMed. *The Lancet. Oncology* **17**(2016 Apr).

100. V, R.*, et al.* Therapeutic Impact of Cytoreductive Surgery and Irradiation of Posterior Fossa Ependymoma in the Molecular Era: A Retrospective Multicohort Analysis - PubMed. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **34**(07/20/2016).

101. R, V.*, et al.* Supratentorial ependymoma in children: to observe or to treat following gross total resection? - PubMed. *Pediatric blood & cancer* **58**(2012 Mar).

102. KW, P.*, et al.* The current consensus on the clinical management of intracranial ependymoma and its distinct molecular variants - PubMed. *Acta neuropathologica* **133**(2017 Jan).

103. R, D.*, et al.* DNA methylation subclasses predict the benefit from gross total tumor resection in IDH-wildtype glioblastoma patients - PubMed. *Neuro-oncology* **25**(02/14/2023).

104. Wijnenga, M.M.J.*, et al.* The impact of surgery in molecularly defined low-grade glioma: an integrated clinical, radiological, and molecular analysis. *Neuro-Oncology* **20**(2018/01/10).

105. Gorzynski, J.E.*, et al.* Ultrarapid Nanopore Genome Sequencing in a Critical Care Setting. *New England Journal of Medicine* **386**(2022).

106. M, L., S, M. & M, S. Real-time selective sequencing using nanopore technology - PubMed. *Nature methods* **13**(2016 Sep).

107. Payne, A.*, et al.* Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nature Biotechnology 2020 39:4* **39**(2020).

108. S, K., Y, F., B, N., W, T. & MC, S. Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED - PubMed. *Nature biotechnology* **39**(2021 Apr).

109. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**(2018).

110. Payne, A.*, et al.* Barcode aware adaptive sampling for GridION and PromethION Oxford Nanopore sequencers. *bioRxiv* (2022).

111. Weilguny, L.*, et al.* Dynamic, adaptive sampling during nanopore sequencing using Bayesian experimental design. *Nature Biotechnology 2023 41:7* **41**(2023).

112. Yuan, D.*, et al.* crossNN: an explainable framework for cross-platform DNA methylation-based classification of cancer. *medRxiv* (2024).

113. Vermeulen, C.*, et al.* Ultra-fast deep-learned CNS tumour classification during surgery. *Nature 2023 622:7984* **622**(2023).

114. Simon, M.*, et al.* Rapid DNA methylation-based classification of pediatric brain tumours from ultrasonic aspirate specimens. *medRxiv* (2023).

115. Afflerbach, A.-K.*, et al.* Nanopore sequencing from formalin-fixed paraffin-embedded specimens for copy-number profiling and methylation-based CNS tumor classification. *Acta Neuropathologica* **147**(2024).

116. Robust methylation-based classification of brain tumours using nanopore sequencing - PubMed. *Neuropathology and applied neurobiology* **49**(2023).

117. Versatile, accessible cross-platform molecular profiling of central nervous system tumors: web-based, prospective multi-center validation. (2024).

118. Patel, A.*, et al.* Rapid-CNS2: rapid comprehensive adaptive nanopore-sequencing of CNS tumors, a proof-of-concept study. *Acta Neuropathologica 2022 143:5* **143**(2022-03-31).

119. Kolmogorov, M.*, et al.* Scalable Nanopore sequencing of human genomes provides a comprehensive view of haplotype-resolved variation and methylation. *Nature Methods 2023 20:10* **20**(2023-09-14).

120. Sahm, F.*, et al.* Next-generation sequencing in routine brain tumor diagnostics enables an integrated diagnosis and identifies actionable targets. *Acta Neuropathologica 2015 131:6* **131**(2015-12-15).

121. Modernizing reference genome assemblies - PubMed. *PLoS biology* **9**(2011 Jul).

122. Payne, A.*, et al.* Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nature Biotechnology 2020 39:4* **39**(2020-11-30).

123. Payne, A.*, et al.* Barcode aware adaptive sampling for GridION and PromethION Oxford Nanopore sequencers. *bioRxiv* (2022-03-30).

124. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**(2018/09/15).

125. Danecek, P.*, et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**(2021/01/29).

126. Wick, R.R.*, et al.* Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial Genomics* **3**(2017/09/14).

127. The UCSC Genome Browser Database: update 2006 - PubMed. *Nucleic acids research* **34**(01/01/2006).

128. Shafin, K.*, et al.* Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nature Methods 2021 18:11* **18**(2021-11-01).

129. Shafin, K.*, et al.* Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nature Methods 2021 18:11* **18**(2021).

130. O'Connell, K.A.*, et al.* Accelerating genomic workflows using NVIDIA Parabricks. *BMC Bioinformatics 2023 24:1* **24**(2023).

131. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**(2010/03/15).

132. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38**(2010/09).

133. A global reference for human genetic variation. *Nature 2015 526:7571* **526**(2015-09-30).

134. Forbes, S.A.*, et al.* The Catalogue of Somatic Mutations in Cancer (COSMIC). *Current Protocols in Human Genetics* **57**(2008/04/01).

135. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**(2011/11/11).

136. CNVpytor: a tool for copy number variation detection and analysis from read depth and allele imbalance in whole-genome sequencing - PubMed. *GigaScience* **10**(11/18/2021).

137. Wright, M.N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* **77**(2017/03/31).

138. Smolka, M.*, et al.* Detection of mosaic and population-level structural variants with Sniffles2. *Nature Biotechnology 2024* (2024-01-02).

139. Heller, D. & Vingron, M. SVIM: structural variant identification using mapped long reads. *Bioinformatics* **35**(2019/09/01).

140. Geoffroy, V.*, et al.* The AnnotSV webserver in 2023: updated visualization and ranking. *Nucleic Acids Research* **51**(2023/07/05).

141. Thorvaldsdóttir, H., Robinson, J.T. & Mesirov, J.P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14**(2013/03/01).

142. Stichel, D.*, et al.* Routine RNA sequencing of formalin-fixed paraffin-embedded specimens in neuropathology diagnostics identifies diagnostically and therapeutically relevant gene fusions. *Acta Neuropathologica 2019 138:5* **138**(2019-07-05).

143. Capper, D.*, et al.* Practical implementation of DNA methylation and copy-number-based CNS tumor diagnostics: the Heidelberg experience. *Acta Neuropathologica* **136**(2018).

144. Capper, D.*, et al.* Practical implementation of DNA methylation and copy-number-based CNS tumor diagnostics: the Heidelberg experience. *Acta Neuropathologica 2018 136:2* **136**(2018-07-02).

145. DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly - PubMed. *Genome research* **24**(2014 Dec).

146. Deacon, S.*, et al.* 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. *medRxiv* (2024-09-11).

147. Versatile, accessible cross-platform molecular profiling of central nervous system tumors: web-based, prospective multi-center validation. (2024-04-10).

148. Patel, A.*, et al.* Rapid-CNS2: rapid comprehensive adaptive nanopore-sequencing of CNS tumors, a proof-of-concept study. *Acta Neuropathologica 2022 143:5* **143**(2022).

149. Lee, H.*, et al.* Detection of TERT Promoter Mutations Using Targeted Next-Generation Sequencing: Overcoming GC Bias through Trial and Error. *Cancer Research and Treatment : Official Journal of Korean Cancer Association* **54**(2022/01).

150. Accuracy and efficiency of germline variant calling pipelines for human genome data - PubMed. *Scientific reports* **10**(11/19/2020).

151. MGMT gene silencing and benefit from temozolomide in glioblastoma - PubMed. *The New England journal of medicine* **352**.

152. Incekara, F.*, et al.* The Association Between the Extent of Glioblastoma Resection and Survival in Light of MGMT Promoter Methylation in 326 Patients With Newly Diagnosed IDH-Wildtype Glioblastoma. *Frontiers in Oncology* **10**(2020).

153. F, G.*, et al.* Surgery for Glioblastoma in Light of Molecular Markers: Impact of Resection and MGMT Promoter Methylation in Newly Diagnosed IDH-1 Wild-Type Glioblastomas - PubMed. *Neurosurgery* **84**(01/01/2019).

154. MGMT methylation analysis of glioblastoma on the Infinium methylation BeadChip identifies two distinct CpG regions associated with gene silencing and outcome, yielding a prediction model for comparisons across datasets, tumor grades, and CIMP-status - PubMed. *Acta neuropathologica* **124**(2012 Oct).

155. High density DNA methylation array is a reliable alternative for PCR-based analysis of the MGMT promoter methylation status in glioblastoma. *Pathology - Research and Practice* **216**(2020/01/01).

156. Accurate and comprehensive evaluation of O6-methylguanine-DNA methyltransferase promoter methylation by nanopore sequencing - PubMed. *Neuropathology and applied neurobiology* **50**(2024 Jun).

157. MGMT gene silencing and benefit from temozolomide in glioblastoma - PubMed. *The New England journal of medicine* **352**(03/10/2005).

158. N2M2 (NOA-20) phase I/II trial of molecularly matched targeted therapies plus radiotherapy in patients with newly diagnosed non-MGMT hypermethylated glioblastoma - PubMed. *Neuro-oncology* **21**(01/01/2019).

159. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma - PubMed. *The New England journal of medicine* **352**(03/10/2005).

160. McAleenan, A.*, et al.* Prognostic value of test(s) for O6-methylguanine–DNA methyltransferase (MGMT) promoter methylation for predicting overall survival in people with glioblastoma treated with temozolomide. *The Cochrane Database of Systematic Reviews* **2021**(2021).

161. Euskirchen, P.*, et al.* Same-day genomic and epigenomic diagnosis of brain tumors using real-time nanopore sequencing. *Acta Neuropathologica 2017 134:5* **134**(2017-06-21).

162. Kuschel, L.P.*, et al.* Neuropathology and Applied Neurobiology | BNS Journal | Wiley Online Library. *Neuropathology and Applied Neurobiology* **49**(2023/02/01).

163. Vermeulen, C.*, et al.* Ultra-fast deep-learned CNS tumour classification during surgery. *Nature 2023 622:7984* **622**(2023-10-11).

164. Djirackor, L.*, et al.* Intraoperative DNA methylation classification of brain tumors impacts neurosurgical strategy. *Neuro-oncology Advances* **3**(Jan-Dec 2021).

165. Takami, H.*, et al.* Distinct patterns of copy number alterations may predict poor outcome in central nervous system germ cell tumors. *Scientific Reports 2023 13:1* **13**(2023-09-21).

166. Maas, S.L.N.*, et al.* Integrated Molecular-Morphologic Meningioma Classification: A Multicenter Retrospective Analysis, Retrospectively and Prospectively Validated. *Journal of Clinical Oncology* **39**(2021-10-07).

167. Sahm, F.*, et al.* cIMPACT-NOW Update 8: Clarifications on molecular risk parameters and recommendations for WHO grading of meningiomas. *Neuro-Oncology* (2024).

168. Raleigh, D.*, et al.* Pan-cancer copy number variant analysis identifies optimized size thresholds and co-occurrence models for individualized risk-stratification. *Research Square* (2024).

169. Sedlazeck, F.J.*, et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods 2018 15:6* **15**(2018-04-30).

170. Jones, D.T.W.*, et al.* Tandem duplication producing a novel oncogenic BRAF fusion gene defines the majority of pilocytic astrocytomas. *Cancer research* **68**(2008/11/11).

171. Sarcomas With CIC-rearrangements Are a Distinct Pathologic Entity With Aggressive Outcome: A Clinicopathologic and Molecular Study of 115 Cases - PubMed. *The American journal of surgical pathology* **41**(2017 Jul).

172. Pediatric-type high-grade neuroepithelial tumors with CIC gene fusion share a common DNA methylation signature - PubMed. *NPJ precision oncology* **7**(03/24/2023).

173. Gambarotti, M.*, et al.* CIC–DUX4 fusion-positive round-cell sarcomas of soft tissue and bone: a single-institution morphological and molecular analysis of seven cases. *Histopathology* **69**(2016/10/01).

174. High prevalence of CIC fusion with double-homeobox (DUX4) transcription factors in EWSR1-negative undifferentiated small blue round cell sarcomas - PubMed. *Genes, chromosomes & cancer* **51**(2012 Mar).

175. Camacho, C.*, et al.* ElasticBLAST: accelerating sequence search via cloud computing. *BMC Bioinformatics 2023 24:1* **24**(2023-03-26).

176. Basic local alignment search tool - PubMed. *Journal of molecular biology* **215**(10/05/1990).

177. EGFRvIII mutations can emerge as late and heterogenous events in glioblastoma development and promote angiogenesis through Src activation - PubMed. *Neuro-oncology* **18**(2016 Dec).

178. An, Z., Aksoy, O., Zheng, T., Fan, Q.-W. & Weiss, W.A. Epidermal growth factor receptor (EGFR) and EGFRvIII in glioblastoma (GBM): signaling pathways and targeted therapies. *Oncogene* **37**(2018/03).

179. Álvarez-Vázquez, A.*, et al.* EGFR amplification and EGFRvIII predict and participate in TAT-Cx43266–283 antitumor response in preclinical glioblastoma models. *Neuro-Oncology* **26**(2024/07/05).

180. Li, A.*, et al.* TGF-β and BMP signaling are associated with the transformation of glioblastoma to gliosarcoma and then osteosarcoma. *Neuro-Oncology Advances* **6**(2024/01/01).

181. Andaloussi-Saghir, K.*, et al.* Secondary gliosarcoma after the treatment of primary glioblastoma multiforme. *North American Journal of Medical Sciences* **3**(2011/11).

182. Yuan, D.*, et al.* crossNN: an explainable framework for cross-platform DNA methylation-based classification of cancer. *medRxiv* (2024-01-23).

183. Yu, F.*, et al.* Heterogeneity and predictors of the effects of AI assistance on radiologists. *Nature Medicine 2024 30:3* **30**(2024-03-19).

184. Schmid, S.*, et al.* VGLL-altered CNS schwannoma is a new tumor entity with distinct DNA-methylation profile and recurrent gene fusions of either VGLL3 or VGLL1. *Brain Pathology* **33**, e13194 (2023).

185. Radke, J.*, et al.* The genomic and transcriptional landscape of primary central nervous system lymphoma. *Nature Communications 2022 13:1* **13**(2022-05-10).

186. Akyurek, N., Uner, A., Benekli, M. & Barista, I. Prognostic significance of MYC, BCL2, and BCL6 rearrangements in patients with diffuse large B-cell lymphoma treated with cyclophosphamide, doxorubicin, vincristine, and prednisone plus rituximab. *Cancer* **118**(2012/09/01).

187. MYC, BCL2, and BCL6 rearrangements in primary central nervous system lymphoma of large B cell type - PubMed. *Annals of hematology* **98**(2019 Jan).

188. S, Z.*, et al.* Long-read sequencing identified intronic repeat expansions in SAMD12 from Chinese pedigrees affected with familial cortical myoclonic tremor with epilepsy - PubMed. *Journal of medical genetics* **56**(2019 Apr).

189. J, S.*, et al.* Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease - PubMed. *Nature genetics* **51**(2019 Aug).

190. Stevanovski, I.*, et al.* Comprehensive genetic diagnosis of tandem repeat expansion disorders with programmable targeted nanopore sequencing. *Science Advances* **8**(2022-03).

191. P, G.*, et al.* Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing - PubMed. *Nature biotechnology* **37**(2019 Dec).

192. Sturm, D.*, et al.* Paediatric and adult glioblastoma: multiform (epi)genomic culprits emerge. *Nature Reviews Cancer 2014 14:2* **14**(2014-01-24).

193. Ceccarelli, M.*, et al.* Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell* **164**(2016/01/28).

194. Zhu, X.*, et al.* The association between telomere length and cancer risk in population studies. *Scientific Reports 2016 6:1* **6**(2016-02-26).

195. Schmidt, T.T.*, et al.* High resolution long-read telomere sequencing reveals dynamic mechanisms in aging and cancer. *Nature Communications 2024 15:1* **15**(2024-06-18).

196. Sholes, S.L.*, et al.* Chromosome-specific telomere lengths and the minimal functional telomere revealed by nanopore sequencing. *Genome Research* **32**(2022/04).

197. Guidelines for Validation of Next-Generation Sequencing-Based Oncology Panels: A Joint Consensus Recommendation of the Association for Molecular Pathology and College of American Pathologists - PubMed. *The Journal of molecular diagnostics : JMD* **19**(2017 May).

198. Pagès-Gallego, M., de Ridder, J., Pagès-Gallego, M. & de Ridder, J. Comprehensive benchmark and architectural analysis of deep learning models for nanopore sequencing basecalling. *Genome Biology 2023 24:1* **24**(2023-04-11).

199. Validation and benchmarking of targeted panel sequencing for cancer genomic profiling - PubMed. *American journal of clinical pathology* **160**(11/02/2023).

200. MacKenzie, M. & Argyropoulos, C. An Introduction to Nanopore Sequencing: Past, Present, and Future Considerations. *Micromachines* **14**(2023/02).

201. Gorzynski, J.E.*, et al.* Ultrarapid Nanopore Genome Sequencing in a Critical Care Setting. *New England Journal of Medicine* **386**(2022-02-17).

202. O'Connell, K.A.*, et al.* Accelerating genomic workflows using NVIDIA Parabricks. *BMC Bioinformatics 2023 24:1* **24**(2023-05-31).

203. Swanton, C. TRACERx EVO: TRACking thoracic Cancer Evolution through therapy (Rx) EVO. (UCL Funder - Cancer Research UK, University College London Hospital, Recruiting, Start date: 13 October 2023, End date: 30 June 2034).

204. Clifford, H. Accelerate Genomic Analysis for Any Sequencer With Parabricks v4.2. (2023).

205. Jain, M.*, et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology 2018 36:4* **36**(2018-01-29).

206. Di Tommaso, P.*, et al.* Nextflow enables reproducible computational workflows. *Nature Biotechnology 2017 35:4* **35**(2017-04-11).

207. Stanojević, D., Lin, D., Sessions, P.F.d. & Šikić, M. Telomere-to-telomere phased genome assembly using error-corrected Simplex nanopore reads. *bioRxiv* (2024-05-21).

208. A Summary of the Inaugural WHO Classification of Pediatric Tumors: Transitioning from the Optical into the Molecular Era - PubMed. *Cancer discovery* **12**(2022).

209. Pidsley, R.*, et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biology 2016 17:1* **17**(2016).

210. Noguera-Castells, A., García-Prieto, C.A., Álvarez-Errico, D. & Esteller, M. Validation of the new EPIC DNA methylation microarray (900K EPIC v2) for high-throughput profiling of the human DNA methylome. *Epigenetics* **18**(2023).

211. Kuschel, L.P.*, et al.* Neuropathology and Applied Neurobiology | BNS Journal | Wiley Online Library. *Neuropathology and Applied Neurobiology* **49**(2023).

212. Euskirchen, P.*, et al.* Same-day genomic and epigenomic diagnosis of brain tumors using real-time nanopore sequencing. *Acta Neuropathologica 2017 134:5* **134**(2017).

213. Djirackor, L.*, et al.* Intraoperative DNA methylation classification of brain tumors impacts neurosurgical strategy. *Neuro-oncology Advances* **3**(2021).

214. Breiman, L. & Breiman, L. Random Forests. *Machine Learning 2001 45:1* **45**(2001).

215. Louppe, G. Understanding Random Forests: From Theory to Practice. (2014).

216. Cutler, A., Cutler, D.R. & Stevens, J.R. Random Forests. in *Ensemble Machine Learning* 157-175 (Springer, 2007).

217. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016).

218. Friedman, J.H. & Friedman, J.H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* **29**(2001).

219. Guan, X.*, et al.* Construction of the XGBoost model for early lung cancer prediction based on metabolic indices. *BMC Medical Informatics and Decision Making* **23**(2023).

220. LeCun, Y.*, et al.* Deep learning. *Nature 2015 521:7553* **521**(2015).

221. SrivastavaNitish, HintonGeoffrey, KrizhevskyAlex, SutskeverIlya & SalakhutdinovRuslan. Dropout. *The Journal of Machine Learning Research* (2014).

222. C., L. The mythos of model interpretability. *Communications of the ACM* **61**(2018-09-26).

223. Tabular data: Deep learning is not all you need. *Information Fusion* **81**(2022).

224. Maros, M.E.*, et al.* Machine learning workflows to estimate class probabilities for precision cancer diagnostics on DNA methylation microarray data. *Nature Protocols 2020 15:2* **15**(2020).

225. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V. & Gulin, A. CatBoost: Unbiased Boosting with Categorical Features. in *Advances in Neural Information Processing Systems*, Vol. 31 (2018).

226. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016/03/09).

227. Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* **28**(2008/11/10).

228. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays - PubMed. *Bioinformatics (Oxford, England)* **30**(05/15/2014).

229. Morrison, J.*, et al.* Evaluation of whole-genome DNA methylation sequencing library preparation protocols. *Epigenetics & Chromatin 2021 14:1* **14**(2021-06-19).

230. Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing - PubMed. *Nature* **510**(06/26/2014).

231. OTP: An automatized system for managing and processing NGS data. *Journal of Biotechnology* **261**(2017/11/10).

232. Vaisvila, R.*, et al.* Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. *Genome Research* **31**(2021/07).

233. Schneider, V.A.*, et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research* **27**(2017-05-01).

234. Lakens, D. & Lakens, D. Equivalence Tests. *Social Psychological and Personality Science* **8**(2017-05-05).

235. Wu, Z.*, et al.* Impact of the methylation classifier and ancillary methods on CNS tumor diagnostics. *Neuro-Oncology* **24**(2022/04/01).

236. Sturm, D.*, et al.* Multiomic neuropathology improves diagnostic accuracy in pediatric neuro-oncology. *Nature Medicine 2023 29:4* **29**(2023-03-16).

237. Bishop, C.M. *Pattern recognition and machine learning*, (Springer New York, NY, 2006).

238. Extracting and composing robust features with denoising autoencoders | Proceedings of the 25th international conference on Machine learning.

239. Katz, S., Santos, V.A.P.M.d., Saccenti, E. & Roshchupkin, G.V. mEthAE: an Explainable AutoEncoder for methylation data. *bioRxiv* (2024-01-19).

240. Titus, A.J., Wilkins, O.M., Bobak, C.A. & Christensen, B.C. Unsupervised deep learning with variational autoencoders applied to breast tumor genome-wide DNA methylation data with biologic feature extraction. *bioRxiv* (2018-11-07).

241. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders | Biocomputing 2018. *Biocomputing 2018* (2018).

242. Owens, A.R.*, et al.* Novel deep learning-based solution for identification of prognostic subgroups in liver cancer (Hepatocellular carcinoma). *BMC Bioinformatics 2021 22:1* **22**(2021-11-24).

243. Choi, J., Chae, H., Choi, J. & Chae, H. methCancer-gen: a DNA methylome dataset generator for user-specified cancer type based on conditional variational autoencoder. *BMC Bioinformatics 2020 21:1* **21**(2020-05-11).

244. Yuen, Z.W.-S.*, et al.* Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing. *Nature Communications 2021 12:1* **12**(2021-06-08).

245. The blood-brain barrier and blood-tumour barrier in brain tumours and metastases - PubMed. *Nature reviews. Cancer* **20**(2020 Jan).

246. Baca, S.C.*, et al.* Liquid biopsy epigenomic profiling for cancer subtyping. *Nature Medicine 2023 29:11* **29**(2023-10-21).

247. Detection of circulating tumor DNA in early- and late-stage human malignancies - PubMed. *Science translational medicine* **6**(02/19/2014).

248. Liquid biopsy in gliomas: A RANO review and proposals for clinical applications - PubMed. *Neuro-oncology* **24**(06/01/2022).

249. Detection and discrimination of intracranial tumors using plasma cell-free DNA methylomes - PubMed. *Nature medicine* **26**(2020 Jul).

250. Iser, F.*, et al.* Cerebrospinal Fluid cfDNA Sequencing for Classification of Central Nervous System Glioma. *Clinical Cancer Research* **30**(2024/07/15).

251. Next-generation sequencing of cerebrospinal fluid for clinical molecular diagnostics in pediatric, adolescent and young adult brain tumor patients - PubMed. *Neuro-oncology* **24**(10/03/2022).

252. Classification of Brain Tumors by Nanopore Sequencing of Cell-Free DNA from Cerebrospinal Fluid - PubMed. *Clinical chemistry* **70**(01/04/2024).

253. Tracking tumour evolution in glioma through liquid biopsies of cerebrospinal fluid - PubMed. *Nature* **565**(2019 Jan).

254. Jeong, Y.*, et al.* MethylBERT: A Transformer-based model for read-level DNA methylation pattern identification and tumour deconvolution. *bioRxiv* (2024-08-29).

255. Elliott, H.R.*, et al.* Characterisation of ethnic differences in DNA methylation between UK-resident South Asians and Europeans. *Clinical Epigenetics 2022 14:1* **14**(2022-10-15).

256. Ethnic differences in cancer risk resulting from genetic variation - PubMed. *Cancer* **86**(12/01/1999).

257. Genetic Misdiagnoses and the Potential for Health Disparities - PubMed. *The New England journal of medicine* **375**(08/18/2016).

258. Chen, C.*, et al.* Sociodemographics and Epigenetic Age Acceleration in Survivors of Childhood Cancer. *JAMA Network Open* **7**(2024/07/01).

259. Jordan, I.K., Lee, K.K., McDonald, J.F. & Mariño-Ramírez, L. Epigenetics and cancer disparities: when nature might be nurture. *Oncoscience* **9**(2022).

260. Zhang, J.-Y.*, et al.* A single-molecule nanopore sequencing platform. *bioRxiv* (2024-08-20).

261. Jurmeister, P.*, et al.* DNA methylation-based classification of sinonasal tumors. *Nature Communications 2022 13:1* **13**(2022-11-28).

262. Koelsche, C.*, et al.* Sarcoma classification by DNA methylation profiling. *Nature Communications 2021 12:1* **12**(2021-01-21).

263. Huskens, N.*, et al.* Closing the Gap Report: A roadmap for equitable access to genomic testing and precision medicine trials for all patients with a brain tumour in the UK. (Tessa Jowell Brain Cancer Mission).

264. Fu, Y.*, et al.* MethPhaser: methylation-based haplotype phasing of human genomes. *bioRxiv* (2023-05-14).

265. Watkins, T.B.K.*, et al.* Refphase: Multi-sample phasing reveals haplotype-specific copy number heterogeneity. *PLOS Computational Biology* **19**(23 Oct 2023).

266. Abbosh, C.*, et al.* Tracking early lung cancer metastatic dissemination in TRACERx using ctDNA. *Nature* **616**(2023/04/04).

267. Bailey, C.*, et al.* Tracking Cancer Evolution through the Disease Course. *Cancer discovery* **11**(2021/04/04).

268. Black, J.R.M., McGranahan, N., Black, J.R.M. & McGranahan, N. Genetic and non-genetic clonal diversity in cancer evolution. *Nature Reviews Cancer 2021 21:6* **21**(2021-03-16).

269. Shmatko, A.*, et al.* Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. *Nature Cancer 2022 3:9* **3**(2022-09-22).

270. JN, K.*, et al.* Pan-cancer image-based detection of clinically actionable genetic alterations - PubMed. *Nature cancer* **1**(2020 Aug).

271. Hoang, D.-T.*, et al.* Prediction of DNA methylation-based tumor types from histopathology in central nervous system tumors with deep learning. *Nature Medicine 2024 30:7* **30**(2024-05-17).

272. Y, F.*, et al.* Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis - PubMed. *Nature cancer* **1**(2020 Aug).

273. Xu, H.*, et al.* A whole-slide foundation model for digital pathology from real-world data. *Nature 2024 630:8015* **630**(2024-05-22).

274. X, W.*, et al.* A pathology foundation model for cancer diagnosis and prognosis prediction - PubMed. *Nature* (09/04/2024).

275. Lu, M.Y.*, et al.* A visual-language foundation model for computational pathology. *Nature Medicine 2024 30:3* **30**(2024-03-19).

276. Lu, M.Y.*, et al.* AI-based pathology predicts origins for cancers of unknown primary. *Nature 2021 594:7861* **594**(2021-05-05).

277. Lu, M.Y.*, et al.* Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering 2021 5:6* **5**(2021-03-01).

278. OSM, E.N.*, et al.* From whole-slide image to biomarker prediction: end-to-end weakly supervised deep learning in computational pathology - PubMed. *Nature protocols* (09/16/2024).

279. Heidelberg Epignostix Gmb, H. Epignostix is Live: New Developments in Molecular Diagnostics.

Appendix A Tables and Figures

A.1 Rapid-CNS^2

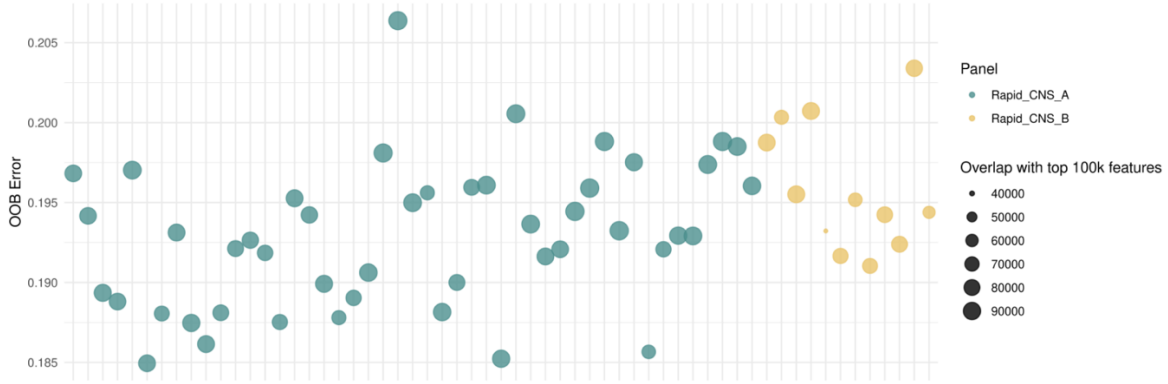
A.1.1 *TERT* promoter mutation concordance

Study ID	*TERTp status*
1	discordant, called in NGS
2	discordant, called in NGS
3	concordant, upstream 1295228
4	concordant, upstream 1295250
5	concordant, upstream 1295228
6	concordant, upstream 1295228
7	concordant, upstream 1295228
8	concordant, upstream 1295250
9	concordant, upstream 1295228
10	discordant, called in NGS
11	discordant, called in NGS
12	concordant, upstream 1295250
13	concordant, upstream 1295228
14	concordant, upstream 1295228
15	concordant, upstream 1295228
17	concordant, upstream 1295228
18	concordant, upstream 1295228
22	concordant, upstream 1295228
29	concordant, upstream 1295228
31	concordant, upstream 1295228

32	concordant, upstream 1295228
34	concordant, upstream 1295228
36	concordant, upstream 1295228
37	concordant, upstream 1295228
45	concordant, upstream 1295228
48	concordant, upstream 1295228
50	concordant, upstream 1295228
51	concordant, upstream 1295250
54	concordant, upstream 1295228
57	concordant, upstream 1295228
60	concordant, upstream 1295228
61	concordant, upstream 1295228
62	concordant, upstream 1295228
65	concordant, upstream 1295228
66	concordant, upstream 1295228
67	concordant, upstream 1295228
68	concordant, upstream 1295228
69	concordant, upstream 1295228
83	concordant, upstream 1295250
86	concordant, upstream 1295228
88	discordant, called in NGS
89	concordant, upstream 1295250
92	concordant, upstream 1295250
105	concordant, upstream 1295228
116	concordant, upstream 1295228
119	concordant, upstream 1295228
120	concordant, upstream 1295228
152	concordant, upstream 1295228
155	concordant, upstream 1295228

156	discordant, called in NGS
171	concordant, upstream 1295228
185	concordant, upstream 1295250
194	concordant, upstream 1295228

A.1.2 Out of the bag (OOB) error for 54 samples



A.1.3 Integrated diagnoses within 30 minutes of sequencing for prospective samples

ID	*Frozen section diagnosis*	*Methylation class (30 min)*	*CNVs (30 min)*	*Integrated diagnosis (30 min)*	*Conventional diagnosis*
170	Meningioma without indication for atypia	MNG		Meningioma WHO grade 1, no high-risk CNVs	Meningioma WHO grade 1
171	Glioblastoma	GBM_RTK_II		Glioblastoma IDH-wildtype WHO grade 4	Glioblastoma IDH-wildtype WHO grade 4
172	High grade glioma	GBM_RTK_II		Glioblastoma IDH-wildtype WHO grade 4	Glioblastoma IDH-wildtype WHO grade 4
173	Cell dense tumor, suspected glioma	EFT_CIC		Cell dense tumor, morphologically suspected glioma, molecularly indicative of CIC-rearranged sarcoma	Round cell sarcomatoid tumor with methylation class "CIC-rearranged sarcoma" but no detection of CIC-fusion by RNA sequencing, NEC
174	Meningioma, highly vasculated	MNG		Meningioma WHO grade 1, no high-risk CNVs	Meningioma WHO grade 1

175	Spindle cell tumor without signs of higher malignancy	MNG		Spindle cell tumor without signs of higher malignancy, chr6 del	Desmoid-type fibromatosis
176	Glioblastoma	GBM_RTK_II		Glioblastoma IDH-wildtype WHO grade 4	Glioblastoma IDH-wildtype WHO grade 4
177	Spindle cell tumor, compatible with schwannoma	SCHW		Schwannoma WHO grade 1	Schwannoma WHO grade 1
178	Isomorphic tumor, compatible with ependymoma	EPN_SPINE		Ependymoma, 22q deleted, several additional CNVs	Spinal ependymoma WHO grade 2
179	Glioma	A_IDH_HG		Astrocytoma IDH-mutant	Astrocytoma IDH-mutant WHO grade 3
180	Malignant glioma	GBM_RTK_II		Glioblastoma IDH-wildtype WHO grade 4	Glioblastoma IDH-wildtype WHO grade 4
181	Meningioma	MNG		Meningioma WHO grade 1, 22q deleted, no high-risk CNVs	Meningioma WHO grade 1
182	High grade glioma	GBM_RTK_II		Glioblastoma IDH-wildtype WHO grade 4	Glioblastoma IDH-wildtype WHO grade 4
183	Meningioma without indication for atypia	MNG		Meningioma WHO grade 1, no high-risk CNVs	Meningioma WHO grade 1
184	Spindle cell tumor without indication for higher malignancy	MELAN	7/10	Glioblastoma IDH-wildtype WHO grade 4	Gliosarcoma IDH-wildtype WHO grade 4
185	Small, blue, round cell tumor, compatible with medulloblastoma	MB_SHH_CHL_AD		Medulloblastoma, SHH-activated, WHO grade 4	Medulloblastoma, SHH-activated and TP53-wildtype, WHO grade 4
186	Spindle cell tumor, fitting to meningioma. No indication of atypia.	MNG		Meningioma WHO grade 1, 22q deleted, no high-risk CNVs	Meningioma WHO grade 1
187	Low grade glial / glioneuronal tumor, compatible with DNET	LGG_DNT		Dysembryoplastic neuroepithelial tumor WHO grade 1	Dysembryoplastic neuroepithelial tumor WHO grade 1
188	Malignant glioma	A_IDH_HG		Astrocytoma IDH-mutant	Astrocytoma IDH-mutant WHO grade 4
189	Glioblastoma	EFT_CIC	7/10	Glioblastoma IDH-wildtype WHO grade 4	Glioblastoma IDH-wildtype WHO grade 4
190	Suspected ependymoma	EPN_SPINE		Ependymoma, 22q deleted	Spinal ependymoma who grade 2
191	Malignant tumor, most likely glioblastoma	GBM_RTK_I		Glioblastoma IDH-wildtype WHO grade 4	Glioblastoma IDH-wildtype WHO grade 4
192	CNS-tissue with elevated cell density and satellitosis, tumor	LGG_DNT	1p/19q	Oligodendroglioma IDH-mutant and 1p/19q-codeleted	Oligodendroglioma IDH-mutant and 1p/19q-codeleted WHO grade 2

	possible, although uncertain				
193	Low grade astrocytic tumor	O_IDH		Oligodendroglioma IDH-mutant and 1p/19q-codeleted	Oligodendroglioma IDH-mutant and 1p/19q-codeleted WHO grade 2
194	Suspected recurrence of the known glioblastoma	GBM_RTK_II		Glioblastoma IDH-wildtype WHO grade 4	Glioblastoma IDH-wildtype WHO grade 4
195	Glioma with no sign of increased malignancy	SUBEPN_PF		Subependymoma	Ependymal glioma with a loss if chr.6q and *TERTp* mutation, NEC
196	Malignant glioma	GBM_RTK_II		Glioblastoma IDH-wildtype WHO grade 4	Glioblastoma IDH-wildtype WHO grade 4
197	Suspected ependymoma	EPN_SPINE		Ependymoma, 22q deleted, several additional CNVs	Spinal ependymoma WHO grade 2
198	Compatible with a glioblastoma	GBM_RTK_I	7/10	Glioblastoma IDH-wildtype WHO grade 4	Glioblastoma IDH-wildtype WHO grade 4
199	Compatible with a central neurocytoma	CN		Central neurocytoma	Central neurocytoma
200	Reactive tissue, compatible with radionecrosis	LYMPHO	no advantage	Reactive tissue, flat CNV	Glioblastoma IDH-wildtype WHO grade 4
201	Glial tumor, compatible with an ependymoma	EPN_SPINE		Ependymoma, 22q deleted, several additional CNVs	Spinal ependymoma WHO grade 2
202	Schwannoma	SCHW		Schwannoma WHO grade 1	Schwannoma WHO grade 1
203	Small blue round cell tumor	MB_G4		Medulloblastoma WHO grade 4, group 4	Medulloblastoma WHO grade 4, group 4
204	Compatible with pilocytic astrocytoma	CONTR_HEMI	5/7 gain	Compatible with pilocytic astrocytoma by histology and CNVs	Pilocytic astrocytoma WHO grade 1
205	Schwannoma	SCHW		Schwannoma WHO grade 1	Schwannoma WHO grade 1
206	Epitheloid tumor	MNG		Meningioma WHO grade 1, multiple chromosomal gains, indicating metaplastic subtype	Meningioma WHO grade 1
207	Meningioma	MNG		Meningioma WHO grade 1, 22q deleted, no high-risk CNVs	Meningioma WHO grade 1
208	Meningioma DD SFT	CONTR_INFLAM	22q del	Meningioma WHO grade 1, 22q deleted, no high-risk CNVs	Meningioma WHO grade 1
209	Malignant glioma	LGG_DNT	segmental 7q gain	Malignant glioma	Glial/glioneuronal tumor, MET fused and VHL mutant, NEC

210	Infiltration zone of a diffuse glioma	O_IDH		Oligodendroglioma IDH-mutant and 1p/19q-codeleted	Oligodendroglioma WHO grade 2
211	Malignant glioma, compatible with glioblastoma	GBM_RTK_II		Glioblastoma IDH-wildtype WHO grade 4	Glioblastoma IDH-wildtype WHO grade 4
212	Compatible with glioblastoma	GBM_RTK_II		Glioblastoma IDH-wildtype WHO grade 4	Glioblastoma IDH-wildtype WHO grade 4
213	Glioblastoma	GBM_RTK_II		Glioblastoma IDH-wildtype WHO grade 4	Glioblastoma IDH-wildtype WHO grade 4
214	Compatible with infiltration zone of a diffuse glioma	A_IDH		Astrocytoma IDH-mutant	Astrocytoma IDH-mutant
215	Malignant glioma	GBM_RTK_II		Glioblastoma IDH-wildtype WHO grade 4	Glioblastoma IDH-wildtype WHO grade 4
216	Infiltration zone of a diffuse glioma	O_IDH		Oligodendroglioma IDH-mutant and 1p/19q-codeleted	Oligodendroglioma WHO grade 2
217	Meningioma	MNG		Meningioma WHO grade 1, no high-risk CNVs	Meningioma WHO grade 1
218	Malignant glioma, compatible with glioblastoma	GBM_RTK_II		Glioblastoma IDH-wildtype WHO grade 4	Glioblastoma IDH-wildtype WHO grade 4
219	Suspected HGAP	DMG_K27	multiple CNVs	Suspected HGAP, compatible CDKN2A/B homozygous deletion	High-grade astrocytoma with piloid features (HGAP)
220	PitNet DD craniopharyngeoma	PITAD_ACTH		Pituitary adenoma / pituitary neuroendocrine tumour (PitNet)	Pituitary adenoma / pituitary neuroendocrine tumour (PitNet)

A.2 MNP-Flex

A.2.1 Abbreviations for MNP-Flex classes

Abbreviation	***Description***
A_IDH_LG	Astrocytoma, IDH-mutant
HGAP	High-grade astrocytoma with piloid features
GTAKA	Glioneuronal tumor with ATRX alteration, kinase fusion and anaplastic features (novel)
ATRT_MYC	Atypical teratoid rhabdoid tumour, MYC activated
ATRT_SHH	Atypical teratoid rhabdoid tumour, SHH activated
ATRT_TYR	Atypical teratoid rhabdoid tumour, Tyrosinase activated

CHGL	Chordoid glioma, PRKCA-mutant
CHORDM	Chordoma
CN	Central neurocytoma
CNS_NB_FOXR2	CNS neuroblastoma, FOXR2-altered
CNS_SARC_DICER	Primary intracranial sarcoma, DICER1-mutant
CPC_PED	Choroid plexus carcinoma, pediatric subtype
CPC_AD	Choroid plexus carcinoma, adult subtype
CPH_ADM	Adamantinomatous craniopharyngioma
CPH_PAP	Papillary craniopharyngioma
CPP_AD	Choroid plexus papilloma, adult subtype
CPP_PED	Choroid plexus papilloma, pediatric subtype
CRINET	Cribriform neuroepithelial tumour
CTRL_BLOOD	Control tissue, blood
DGONC	Diffuse glioneuronal tumour with oligodendroglioma-like features and nuclear clusters
DIG_DIA	Desmoplastic infantile ganglioglioma / desmoplastic infantile astrocytoma
DLGNT_1	Diffuse leptomeningeal glioneuronal tumour, subclass 1
DLGNT_2	Diffuse leptomeningeal glioneuronal tumour, subclass 2
DMG_K27	Diffuse midline glioma, H3 K27-altered, subtype H3 K27-mutant or EZHIP expressing
DNET	Dysembryoplastic neuroepithelial tumour
CNS_SARC_CIC	CIC-rearranged sarcoma
ONB	Olfactory neuroblastoma
SNUC_IDH2	Sinonasal undifferentiated carcinoma, IDH2-mutant
CNS_BCOR_FUS	CNS tumour with EP300:BCOR(L1) fusion
EPN_MPE	Myxopapillary ependymoma
EPN_PF_SE	Posterior fossa subependymoma
EPN_SPINE	Spinal ependymoma
EPN_SPINE_MYCN	Spinal ependymoma, MYCN-amplified
EPN_SPINE_SE_B	Spinal subependymoma [subtype B]
EPN_SPINE_SE_A	Spinal subependymoma [subtype A]
NET_PLAGL1_FUS	Neuroepithelial tumour, PLAGL1-fused
EPN_ST_SE	Supratentorial subependymoma
EPN_YAP	Supratentorial ependymoma, YAP1-fused
ETMR_C19MC	Embryonal tumour with multilayered rosettes, C19MC altered
ETMR_Atyp	Embryonal tumour with multilayered rosettes, atypical
EVNCYT	Extraventricular neurocytoma
EWS	Ewing sarcoma
GBM_CBM	

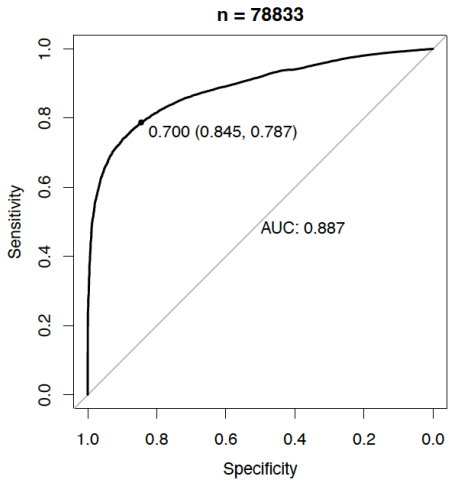
DHG_G34	Diffuse hemispheric glioma, H3 G34-mutant
A_IDH_HG	Astrocytoma, IDH-mutant
INFLAM_ENV	Inflammatory microenvironment
GBM_MES_TYP	Glioblastoma, IDH-wildtype, [typical mesenchymal type]
GBM_MES_ATYP	Glioblastoma, IDH-wildtype, [atypical mesenchymal type]
pedHGG_MYCN	Diffuse paediatric-type high grade glioma, MYCN subtype
pedHGG_RTK1A	Diffuse paediatric-type high grade glioma, RTK1 subtype, subclass A (novel)
pedHGG_RTK1B	Diffuse paediatric-type high grade glioma, RTK1 subtype, subclass B (novel)
pedHGG_RTK1C	Diffuse paediatric-type high grade glioma, RTK1 subtype, subclass C (novel)
pedHGG_RTK2A	Diffuse paediatric-type high grade glioma, RTK2 subtype, subclass A (novel)
pedHGG_RTK2B	Diffuse paediatric-type high grade glioma, RTK2 subtype, subclass B (novel)
GBM_RTK1	Glioblastoma, IDH-wildtype, RTK1 subtype
GBM_RTK2	Glioblastoma, IDH-wildtype, RTK2 subtype
DMG_EGFR	Diffuse midline glioma, H3 K27-altered, subtype EGFR-altered
GCT_GERM_A	Germinoma, subtype KIT wildtype (novel)
GCT_GERM_KIT	Germinoma, subtype KIT mutant (novel)
GCT_TERA	Teratoma
GCT_YOLKSAC	Yolk sac tumour
GG	Ganglioglioma
CTRL_REACTIVE	Control tissue, reactive tumour microenvironment
GNT_A	Diffuse glioneuronal tumour, subtype A
pedHGG_A	Diffuse paediatric-type high grade glioma, H3 wildtype and IDH wild type, Subtype A
pedHGG_B	Diffuse paediatric-type high grade glioma, H3 wildtype and IDH wild type, Subtype B
CNS_BCOR_ITD	CNS tumour with BCOR internal tandem duplication
NET_CXXC5	Neuroepithelial tumour, MN1:CXXC5-fused
ABM_MN1	Astroblastoma, MN1-altered, MN1:BEND2-fused
GBM_PNC	Glioblastoma, IDH-wildtype, with primitive neuronal component
HGG_B	Adult-type diffuse high grade glioma, IDH-wildtype, subtype B
HGG_E	Adult-type diffuse high grade glioma, IDH-wildtype, subtype E
HGG_F	Adult-type diffuse high grade glioma, IDH-wildtype, subtype F
NET_PATZ1	Neuroepithelial tumour with PATZ1 fusion
ET_PLAG	CNS Embryonal tumour with PLAG-family amplification
HMB	Haemangioblastoma
SFT_HMPC	Solitary fibrous tumour / haemangiopericytoma
IHG	Infant-type hemispheric glioma
IO_MEPL	Intraocular medulloepithelioma
LCH	Langerhans cell histiocytosis

AG_MYB	Angiocentric glioma, MYB/MYBL1-altered
LGG_MYB_B	Diffuse astrocytoma, MYB or MYBL1-altered, subtype B [infratentorial] (novel)
LGG_MYB_C	Diffuse astrocytoma, MYB or MYBL1-altered, subtype C [isomorphic] (novel)
LGG_MYB_D	Diffuse astrocytoma, MYB or MYBL1-altered, subtype D (novel)
LIPN	Liponeurocytoma
ET_BRD4_LEUTX	CNS embryonal tumour with BRD4:LEUTX fusion
MB_MYO	Medullomyoblastoma
MB_SHH_1	Medulloblastoma, SHH-activated, subtype 1
MB_SHH_2	Medulloblastoma, SHH-activated, subtype 2
MB_SHH_3	Medulloblastoma, SHH-activated, subtype 3
MB_SHH_4	Medulloblastoma, SHH-activated, subtype 4
MB_SHH_IDH	Medulloblastoma, SHH-activated, IDH-mutant
MB_WNT	Medulloblastoma, WNT activated
MB_G34_I	Medulloblastoma Group 3, subclass I
MB_G34_II	Medulloblastoma Group 3, subclass II
MB_G34_III	Medulloblastoma Group 3, subclass III
MB_G34_IV	Medulloblastoma Group 3, subclass IV
MB_G34_V	Medulloblastoma Group 4, subclass V
MB_G34_VI	Medulloblastoma Group 4, subclass VI
MB_G34_VII	Medulloblastoma Group 4, subclass VII
MB_G34_VIII	Medulloblastoma Group 4, subclass VIII
MMNST	Malignant melanotic nerve sheath tumour
MELN	Melanocytoma
MET_MEL	Melanoma [metastatic]
MNG_BEN_1	Meningioma, subclass benign 1
MNG_BEN_2	Meningioma, subclass benign 2
MNG_BEN_3	Meningioma, subclass benign 3
MNG_SMARCE1	Meningioma, SMARCE1-altered
MNG_INT_A	Meningioma, subclass intermediate A
MNG_INT_B	Meningioma, subclass intermediate B
MNG_MAL	Meningioma, malignant
MPNST_TYP	Malignant peripheral nerve sheath tumour [typical type]
MPNST_ATYP	Malignant peripheral nerve sheath tumour [spinal or atypical type]
MYXGNT	Myxoid glioneuronal tumour, PDGFRA-mutant
NB_MYCN	Neuroblastoma, MYCN subtype
NB_TMM_NEG	Neuroblastoma, subtype TMM negative
NB_TMM_POS	Neuroblastoma, subtype ALT/TERT TMM positive

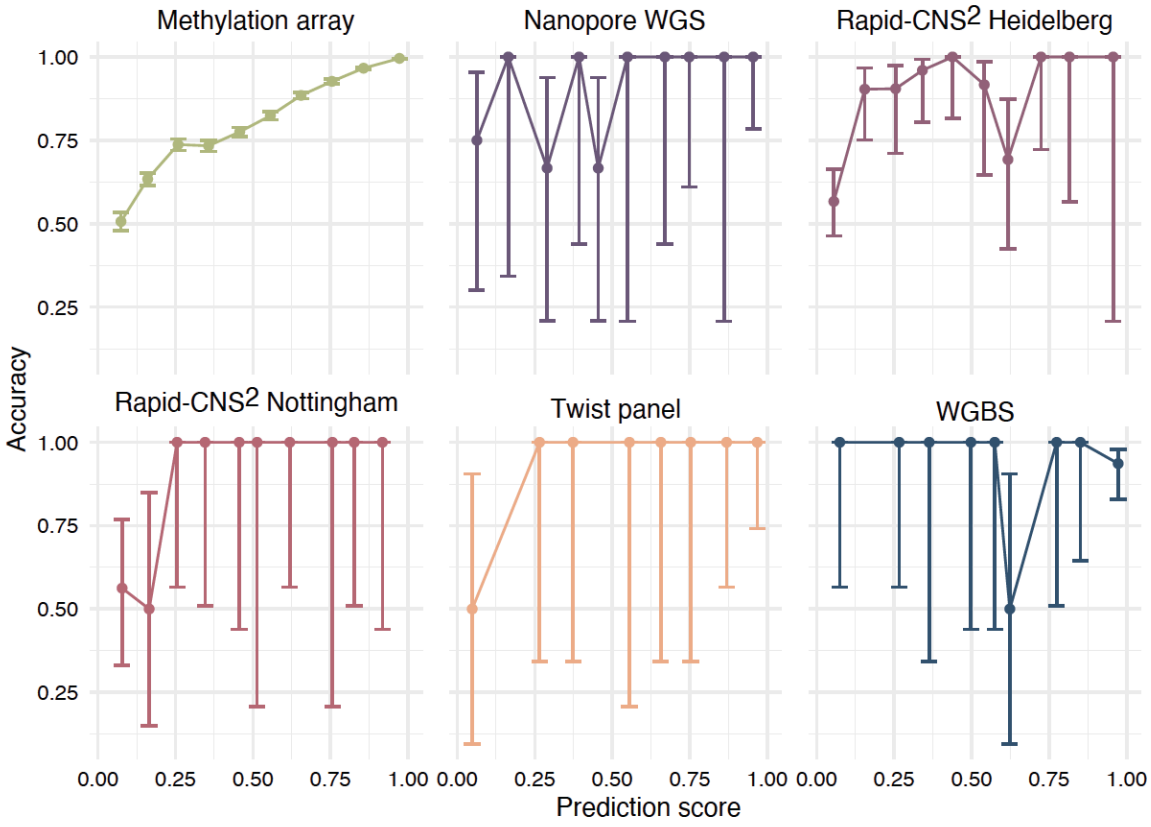
NFIB_PLEX	Plexiform neurofibroma
CTRL_ADENOPIT	Control tissue, pituitary gland (anterior lobe)
CTRL_CBM	Control tissue, cerebellar hemisphere
CTRL_CORPCAL	Control tissue, white matter (corpus callosum)
CTRL_HEMI	Control tissue, hemispheric cortex
CTRL_HYPOTHAL	Control tissue, hypothalamus
CTRL_OPTIC	Control tissue, optic pathway
CTRL_PIN	Control tissue, pineal gland
CTRL_PONS	Control tissue, pons
O_IDH	Oligodendroglioma, IDH-mutant and 1p/19q-codeleted
OLIGOSARC_IDH	Oligosarcoma, IDH-mutant
PA_CORT	Supratentorial pilocytic astrocytoma
PA_INF	Infratentorial pilocytic astrocytoma
PA_INF_FGFR	Infratentorial pilocytic astrocytoma, FGFR1-altered
PA_MID	Supratentorial midline pilocytic astrocytoma
PB_GRP1A	Pineoblastoma, miRNA pathway altered, group 1A
PB_GRP1B	Pineoblastoma, miRNA pathway altered, group 1B
PB_GRP2	Pineoblastoma, miRNA pathway altered, group 2
PB_FOXR2	Pineoblastoma, MYC/FOXR2-activated
DLBCL	Primary diffuse large B cell lymphoma of the CNS
PLASMACYT	Plasmacytoma of the CNS
EPN_PFA_1A	Posterior fossa group A (PFA) ependymoma, subclass 1a
EPN_PFA_1B	Posterior fossa group A (PFA) ependymoma, subclass 1b
EPN_PFA_1C	Posterior fossa group A (PFA) ependymoma, subclass 1c
EPN_PFA_1D	Posterior fossa group A (PFA) ependymoma, subclass 1d
EPN_PFA_1E	Posterior fossa group A (PFA) ependymoma, subclass 1e
EPN_PFA_1F	Posterior fossa group A (PFA) ependymoma, subclass 1f
EPN_PFA_2A	Posterior fossa group A (PFA) ependymoma, subclass 2a
EPN_PFA_2B	Posterior fossa group A (PFA) ependymoma, subclass 2b
EPN_PFA_2C	Posterior fossa group A (PFA) ependymoma, subclass 2c
EPN_PFB_1	Posterior fossa group B (PFB) ependymoma, subclass 1
EPN_PFB_2	Posterior fossa group B (PFB) ependymoma, subclass 2
EPN_PFB_3	Posterior fossa group B (PFB) ependymoma, subclass 3
EPN_PFB_4	Posterior fossa group B (PFB) ependymoma, subclass 4
EPN_PFB_5	Posterior fossa group B (PFB) ependymoma, subclass 5
CAUDEQU_NET	Cauda equina neuroendocrine tumour [paraganglioma], subtype non-CIMP
PGNT	Papillary glioneuronal tumour

PIN_CYT	Pineocytoma
PIN_RB	Pineal retinoblastoma
PITAD_ACTH	Pituitary adenoma, ACTH-producing
PITAD_GON	Pituitary adenoma, gonadotrophin-producing
PITAD_PRL	Pituitary adenoma, prolactin-producing
PITAD_STH_DENSE1	Pituitary adenoma, STH-producing, subclass densely granulated A
PITAD_STH_DENSE2	Pituitary adenoma, STH-producing, subclass densely granulated B
PITAD_STH_SPARSE	Pituitary adenoma, STH-producing, subclass sparsely granulated
PITAD_TSH	Pituitary adenoma, TSH-producing
PITUI	Pituicytoma, granular cell tumour of the sellar region, and spindle-cell oncocytoma
PLNTY	Polymorphous low-grade neuroepithelial tumour of the young
PPTID_A	Pineal parenchymal tumour of intermediate differentiation, subclass A
PPTID_B	Pineal parenchymal tumour of intermediate differentiation, subclass B
PTPR_A	Papillary tumour of the pineal region, subclass A
PTPR_B	Papillary tumour of the pineal region, subclass B
PXA	Pleomorphic xanthoastrocytoma
RB	Retinoblastoma
RB_MYCN	Retinoblastoma, MYCN-activated
EPN_ST_ZFTA_FUS_C	Supratentorial ependymoma, ZFTA fusion-positive, subclass C
EPN_ST_ZFTA_FUS_D	Supratentorial ependymoma, ZFTA fusion-positive, subclass D
EPN_ST_ZFTA_FUS_E	Supratentorial ependymoma, ZFTA fusion-positive, subclass E
RGNT	Rosette-forming glioneuronal tumour
ARMS	Rhabdomyosarcoma, alveolar subtype
ERMS	Rhabdomyosarcoma, embryonal subtype
RMS_MYOD1	Rhabdomyosarcoma, MYOD1-mutant
SCHW	Schwannoma
SEGA	Subependymal giant cell astrocytoma
EPN_ST_ZFTA_RELA_A	Supratentorial ependymoma, ZFTA fusion-positive, subtype ZFTA-RELA fused, subclass A
EPN_ST_ZFTA_RELA_B	Supratentorial ependymoma, ZFTA fusion-positive, subtype ZFTA-RELA fused, subclass B
CNS_SCHW_VGLL	CNS Schwannoma, VGLL-fused

A.2.2 ROC-AUC for all methylation array samples with MNP-RF score ≥ 0.7



A.2.3 MNP-Flex accuracy over binned prediction scores for each technology



*Bars indicate 95% confidence interval

A.2.4 Nanopore WGS samples for MNP-Flex

Sample ID	*MNP-Flex subclass*	*MNP-Flex subclass score*	*Number of missing sites*	*Other evidence*
ONT_WGS_1	Supratentorial pilocytic astrocytoma	0.66047812	7	Low grade neuroepithelial tumor, compatible with pilocytic astrocytoma (supratentorial)
ONT_WGS_2	Glioblastoma, IDH-wildtype, RTK2 subtype	0.96216446	3	Glioblastoma, IDH-wildtype; TERT promoter, NF1 mutation

ONT_WGS_3	Glioblastoma, IDH-wildtype, RTK1 subtype	0.97804058	13	Glioblastoma, IDH-wildtype
ONT_WGS_4	Meningioma, subclass benign 1	0.91916049	6	Atypical meningioma (brain invasion)
ONT_WGS_5	Diffuse midline glioma, H3 K27-altered, subtype H3 K27-mutant or EZHIP expressing	0.69325751	11	Diffuse midline glioma, H3-3A p.K28M, TP53,ATRX mutations
ONT_WGS_6	Glioblastoma, IDH-wildtype, [typical mesenchymal type]	0.72031081	10	Glioblastoma, IDH-wildtype; TERT promoter, BRAF p.G466V mutation
ONT_WGS_7	High-grade astrocytoma with piloid features	0.74914902	19	Low grade astrocytic tumor, compatible with pilocytic astrocytoma; Slightly increased proliferative activity
ONT_WGS_8	Medulloblastoma, SHH-activated, subtype 1	0.95206517	12	Medulloblastoma, MBEN, SHH
ONT_WGS_9	Supratentorial pilocytic astrocytoma	0.44090471	6	PA supratentorial
ONT_WGS_10	Infratentorial pilocytic astrocytoma	0.92838794	11	PA infratentorial; BRAF:KIAA1549 fusion
ONT_WGS_11	Glioblastoma, IDH-wildtype, [atypical mesenchymal type]	0.07266787	5	High grade astrocytoma with features of GBM; TERT promoter mutation
ONT_WGS_12	Astrocytoma, IDH-mutant, high grade	0.99313051	10	Astrocytoma (CNS WHO grade 3), IDH-mutant
ONT_WGS_13	Oligodendroglioma, IDH-mutant and 1p/19q-codeleted	0.96982682	5	Oligodendroglioma CNS WHO grade 3; 1p/19q deletion
ONT_WGS_14	Glioblastoma, IDH-wildtype, [atypical mesenchymal type]	0.42014417	6	Gliosarcoma; TERT mutation; mesenchymal type
ONT_WGS_15	Glioblastoma, IDH-wildtype, [typical mesenchymal type]	0.35345697	8	Glioblastoma, IDH-wildtype; TERT promoter mutation
ONT_WGS_16	Glioblastoma, IDH-wildtype, [typical mesenchymal type]	0.86063206	13	Glioblastoma, IDH-wildtype; TERT promoter mutation
ONT_WGS_17	Glioblastoma, IDH-wildtype, RTK2 subtype	0.77491069	595	Glioblastoma, IDH-wildtype; TERT promoter mutation; EGFR amplification

ONT_WGS_18	Glioblastoma, IDH-wildtype, RTK1 subtype	0.65486097	10	Glioblastoma, IDH-wildtype; TERT promoter mutation; EGFR amplification
ONT_WGS_19	Astrocytoma, IDH-mutant, high grade	0.96108365	4	Astrocytoma (CNS WHO grade 3), IDH-mutant
ONT_WGS_20	Astrocytoma, IDH-mutant, high grade	0.92733258	10	Astrocytoma (CNS WHO grade 3), IDH-mutant
ONT_WGS_21	Glioblastoma, IDH-wildtype, [typical mesenchymal type]	0.05869766	8	Glioblastoma, IDH-wildtype; TERT promoter mutation
ONT_WGS_22	Glioblastoma, IDH-wildtype, RTK2 subtype	0.54883164	15	Glioblastoma, IDH-wildtype; TERT promoter mutation; EGFR amplification and mutation
ONT_WGS_23	Supratentorial pilocytic astrocytoma	0.30169761	12	Ganglioglioma; BRAF p.V600E
ONT_WGS_24	Oligodendroglioma, IDH-mutant and 1p/19q-codeleted	0.98866129	9	Oligodendroglioma CNS WHO grade 3; 1p/19q deletion
ONT_WGS_25	Astrocytoma, IDH-mutant, high grade	0.98384124	8	Astrocytoma (CNS WHO grade 3), IDH-mutant
ONT_WGS_26	Infratentorial pilocytic astrocytoma	0.48420542	7	Supratentorial PA
ONT_WGS_27	Supratentorial ependymoma, ZFTA fusion-positive, subtype ZFTA-RELA fused, subclass A	0.96517265	7	Ependymoma ZFTA::RELA fusion
ONT_WGS_28	Astrocytoma, IDH-mutant, high grade	0.71385247	7	Astrocytoma (CNS WHO grade 3), IDH-mutant
ONT_WGS_29	Glioblastoma, IDH-wildtype, [typical mesenchymal type]	0.149838	8	Glioblastoma, IDH-wildtype; TERT promoter mutation
ONT_WGS_30	Melanoma [metastatic]	0.04253213	5	Melanoma
ONT_WGS_31	Supratentorial pilocytic astrocytoma	0.07866751	7	Ganglioglioma; BRAF p.V600E
ONT_WGS_32	Astrocytoma, IDH-mutant, high grade	0.7783125	11	Astrocytoma (CNS WHO grade 3), IDH-mutant
ONT_WGS_33	Oligodendroglioma, IDH-mutant and 1p/19q-codeleted	0.29146042	12	Oligodendroglioma CNS WHO grade 3; 1p/19q deletion

ONT_WGS_34	Glioblastoma, IDH-wildtype, RTK2 subtype	0.27574193	6	Glioblastoma, IDH-wildtype; TERT promoter mutation
ONT_WGS_35	Oligodendroglioma, IDH-mutant and 1p/19q-codeleted	0.90576172	23	Oligodendroglioma CNS WHO grade 3; 1p/19q deletion
ONT_WGS_36	Astrocytoma, IDH-mutant, high grade	0.92524707	12	Astrocytoma (CNS WHO grade 3), IDH-mutant
ONT_WGS_37	Glioblastoma, IDH-wildtype, RTK2 subtype	0.75149095	17	Glioblastoma, IDH-wildtype; TERT promoter mutation
ONT_WGS_38	Glioblastoma, IDH-wildtype, RTK2 subtype	0.43869066	4	Glioblastoma, IDH-wildtype; TERT promoter mutation
ONT_WGS_39	Glioblastoma, IDH-wildtype, [typical mesenchymal type]	0.40496102	4	Glioblastoma, IDH-wildtype; TERT promoter mutation
ONT_WGS_40	Glioblastoma, IDH-wildtype, [typical mesenchymal type]	0.17933275	9	Glioblastoma, IDH-wildtype; TERT promoter mutation

A.2.5 Twist methylation panel samples for MNP-Flex

Sample ID	*MNP-Flex subclass*	*MNP-Flex subclass score*	*Number of missing sites*	*MNP-RF subclass*	*Concordance*
Twist_1	Medulloblastoma Group 4, subclass VIII	0.99945432	850	Medulloblastoma Group 4, subclass VIII	correct, subclass level
Twist_2	Oligodendroglioma, IDH-mutant and 1p/19q-codeleted	0.96556652	1038	not classifiable, low array scores	correct, inferred subclass level
Twist_3	Oligodendroglioma, IDH-mutant and 1p/19q-codeleted	0.03357558	20576	Oligodendroglioma, IDH-mutant and 1p/19q-codeleted	correct, subclass level
Twist_4	Medulloblastoma, WNT activated	0.36605892	1057	Medulloblastoma, WNT activated	correct, subclass level
Twist_5	Glioblastoma, IDH-wildtype, RTK1 subtype	0.28064326	924	Glioblastoma, IDH-wildtype, RTK1 subtype	correct, subclass level
Twist_6	Diffuse paediatric-type high grade glioma, MYCN subtype	0.06243469	864	Diffuse paediatric-type high grade glioma, MYCN subtype	correct, subclass level
Twist_7	Rhabdomyosarcoma, alveolar subtype	0.95710695	871	no array	correct, inferred subclass level

Twist_8	Rhabdomyosarcoma, alveolar subtype	0.63181114	986	no array	correct, inferred subclass level
Twist_9	Glioblastoma, IDH-wildtype, RTK2 subtype	0.3801648	718	Glioblastoma, IDH-wildtype, RTK2 subtype	correct, subclass level
Twist_10	Glioblastoma, IDH-wildtype, RTK2 subtype	0.24995735	1302	Glioblastoma, IDH-wildtype, RTK1 subtype	correct, family level
Twist_11	Glioblastoma, IDH-wildtype, RTK1 subtype	0.88813639	756	Glioblastoma, IDH-wildtype, RTK1 subtype	correct, subclass level
Twist_12	Glioblastoma, IDH-wildtype, RTK1 subtype	0.55524468	1128	Glioblastoma, IDH-wildtype, RTK1 subtype	correct, subclass level
Twist_13	Glioblastoma, IDH-wildtype, [typical mesenchymal type]	0.86691952	1130	Glioblastoma, IDH-wildtype, [typical mesenchymal type]	correct, subclass level
Twist_14	Glioblastoma, IDH-wildtype, RTK1 subtype	0.775702	1225	Glioblastoma, IDH-wildtype, RTK1 subtype	correct, subclass level
Twist_15	Glioblastoma, IDH-wildtype, RTK1 subtype	0.68152976	1141	Glioblastoma, IDH-wildtype, RTK1 subtype	correct, subclass level
Twist_16	Medulloblastoma, SHH-activated, subtype 3	0.85646999	1096	Medulloblastoma, SHH-activated, subtype 3	correct, subclass level
Twist_17	Medulloblastoma, WNT activated	0.98816001	1138	Medulloblastoma, WNT activated	correct, subclass level
Twist_18	Medulloblastoma, SHH-activated, subtype 3	0.96852273	1216	Medulloblastoma, SHH-activated, subtype 3	correct, subclass level
Twist_19	Medulloblastoma, SHH-activated, subtype 3	0.84640223	1190	Medulloblastoma, SHH-activated, subtype 3	correct, subclass level
Twist_20	Astrocytoma, IDH-mutant, high grade	0.92211211	1165	Astrocytoma, IDH-mutant, high grade	correct, subclass level
Twist_21	Meningioma, subclass benign 1	0.98504096	3437	Meningioma, subclass benign 1	correct, subclass level
Twist_22	Meningioma, subclass benign 2	0.73003393	3527	Meningioma, subclass benign 2	correct, subclass level
Twist_23	Meningioma, subclass benign 1	0.97072429	4864	Meningioma, subclass benign 1	correct, subclass level
Twist_24	Oligodendroglioma, IDH-mutant and 1p/19q-codeleted	0.9690963	3360	Oligodendroglioma, IDH-mutant and 1p/19q-codeleted	correct, subclass level
Twist_25	Meningioma, subclass benign 2	0.93222749	6724	Meningioma, subclass benign 2	correct, subclass level
Twist_26	Meningioma, subclass benign 2	0.88537419	6668	Meningioma, subclass benign 2	correct, subclass level

Twist_27	Glioblastoma, IDH-wildtype, RTK2 subtype	0.98383969	3302	Glioblastoma, IDH-wildtype, RTK2 subtype	correct, subclass level