Utilising emerging technologies for the molecular diagnosis of genomic pathology

Laura A. Crinnion

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

The University of Leeds

School of Medicine

June 2025

I confirm that the work submitted is my own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgements

In what eventually became an almost ten-year journey to completing my part-time PhD studies, the world around me changed in more ways than one. From a professional standpoint, the reorganisation of genetic testing in the NHS in England brought challenges in terms of new local structures and ways of working, but also provided opportunities in supporting my research, with new equipment and infrastructure enabling scientific developments which would not otherwise have been possible. From a personal perspective, the arrival of my two sons along the way brought enormous joy, but also made the work/study/life balance a little bit more complicated! Throwing in a global pandemic around the mid-way point, the road to completing this thesis was far from simple, and therefore there are a number of people without whom this work would not have been possible.

First, my supervision team; thank you to Professor David Bonthron for providing encouragement and guidance both professionally and academically. I am also extremely grateful to Sir Alex Markham for helping to fund this work. A special thank you to Dr Ian Carr for seeing the potential in me and enabling me to have this opportunity. I am also indebted to Dr Christopher Watson, who has helped to steer the research in the right direction, has developed many of the bioinformatics pipelines described here, and from whom I have learned so much in the years we have been working together.

Special thanks to the team on both sides of the Translational Genomics Facility, who have supported me along the way. On the University of Leeds side, Carolina and Morag for their support and encouragement, often over much-needed coffee and cake! On the NHS side there are too many names to mention, but in particular Jim for picking up the slack in recent months as I reached the final hurdle, and Jak for contributing so much to the long-read work over the last couple of years. Thank you to the Gastrohepatology Team in Sheffield for the clinical support on the TTC37 case. Thank you to the wider NHS team both for allowing me to do this work, and for continuing to support translational research in these challenging times, in the hope that we can still achieve better outcomes for our patients. I am also grateful to Jonathan Coxhead at Newcastle University Genomics Core Facility for allowing me to utilise the 10X Chromium instrument for this work and Raf Hussein for providing training and guidance.

Finally, a huge thank you to all of my friends and family who have spurred me on along the way. All of your many words of encouragement have kept me going. My sons Sam and Ned, who are largely oblivious to this, but I hope will one day be proud of this achievement. And Dave, who is the best partner and father to our boys, and I could not have done this without him. I'm eternally grateful to Mam & Dad, and Annie & Kenny, for supporting us as a family and enabling me to continue my career.

Sadly, since the beginning of this journey, there are family and friends who are no longer with us, but who have shaped who I am and inspired me in so many ways. In losing them I have been reminded of the vital importance of the NHS and the research which underpins it, and this thesis is dedicated to them.

Abstract

Over the last century, the advancement of understanding of the human genome has been profound, and with that, the development of diagnostic genomic testing has been expansive. Next Generation Sequencing technology has enabled the sequencing of the entire human genome in less than 24 hours, something that was unthinkable merely twenty years ago. The field continues to evolve and regularly presents new opportunities to expand the screening and diagnostic capability of genomics within the National Health Service.

While the use of NGS is now common practice throughout the NHS, there remain challenges which this technology has yet to overcome. Elements of the genome remain refractory to analysis, hampered by limitations of both the assay design and the analysis methods used to interpret the data that is generated. Here we evaluate novel technologies which have the potential to overcome these obstacles and contribute to the expansion of the genomic testing repertoire.

Single cell genome analysis is an emerging area of diagnostics; however, the sample input requirements for NGS dictate that cells must undergo a whole genome amplification step to achieve a quantity of DNA sufficient for NGS analysis. Here we evaluate the methods available to perform this and explore the potential of adapting an in-house copy number screening method for use with single cells.

Gene panel screening is standard practice within genomic diagnostics. There are, however, many genes where variant detection and classification is impeded by the presence of highly homologous pseudogene regions. Accurate distinction between gene and pseudogene is not possible with current hybrid capture preparation techniques. As a result, we assess new methods of library preparation that allow short-read sequencing data from single molecules to be linked, enabling contiguous analysis of multi-kilobase regions of the genome.

A further limitation of gene panel screening is that typically reagents are designed to capture only the exons and adjacent canonical splice sites of each target gene. While short read whole genome sequencing is now becoming more commonplace, allowing examination of deeper intronic variation, the technology does not allow the phasing of recessive variants which is sometimes necessary to allow classification of the mutation, meaning testing needs to be expanded beyond the proband to other family members. Here we utilise long read whole genome sequencing for patients where a single recessive variant has been detected by standard of care, to identify the second mutant disease-causing allele.

With the work presented here we outline the difficulties associated with implementing novel genomic technologies, ranging from minor optimisations to those larger obstacles which can go on to prevent long-term implementation of the assay. Despite these challenges, the results achieved demonstrate that although whole genome sequencing in the NHS is becoming more standardised, there remains utility in exploring alternative methodologies which can push the boundaries of variant detection and yield diagnoses for NHS patients that are not currently possible by standard of care.

Table of Contents

List of Tables

List of Figures

Abbreviations

aCGH	array comparative genomic hybridisation
ART	Assisted Reproductive Technology
bp	Base pair
CNVseq	Copy Number Variation sequencing
CRISPR	Clustered Regularly Interspersed Short Palindromic Repeats
CVS	Chorionic villus sample
DMSO	Dimethyl Sulfoxide
DNA	Deoxyribonucleic acid
ddNTP	Dideoxyribonucleotide
dNTP	Deoxyribonucleotide
DOP-PCR	Degenerate Oligonucleotide Primed PCR
EDTA	Ethylenediaminetetraacetic acid
FISH	Fluorescence In-Situ Hybridisation
GEM	Gel bead-in emulsion
GMSA	Genomic Medicine Service Alliance
GLH	Genomics Laboratory Hubs
GIAB	Genome In A Bottle
HNPCC	Hereditary Non-Polyposis Colorectal Cancer
IGV	Integrative Genomics Viewer
ILS	Illumina Laboratory Services
IUGR	Intrauterine growth restriction
LR-PCR	Long-range PCR
MALBAC	Multiple Annealing and Looping-Based Amplification Cycles
MDA	Multiple displacement amplification
mg	milligram
MMR	mismatch repair
NEB	New England Biolabs
ng	Nanogram

NGS	Next Generation Sequencing
nt	Nucleotide
OMIM	Online Mendelian Inheritance in Man database
ONT	Oxford Nanopore Technologies
PCR	Polymerase Chain Reaction
pg	Picogram
PGD	Preimplantation Genetic Diagnosis
PGS	Preimplantation Genetic Screening
RCF	Relative Centrifugal Force
RIN^e	Ribonucleic acid integrity number equivalent
RNA	Ribonucleic acid
RNP	Ribonucleoprotein
smMIP	single molecule molecular inversion probe
SMRT	Single molecule real-time
SNP	Single Nucleotide Polymorphism
TELL-Seq	Transposase Enzyme Linked Long-Read Sequencing
TAE	Tris-Acetate-EDTA
TE	Tris-Ethylenediaminetetraacetic acid (EDTA)
THES	Tricho-hepato-enteric syndrome
TTC37	Tetratricopeptide repeat domain 37
WES	Whole Exome Sequencing
WGA	Whole Genome Amplification
WGS	Whole Genome Sequencing

1. Introduction

1.1 General Overview

The advent and ubiquitous adoption of next-generation (also termed massively parallel) DNA sequencing has, over the past decade, transformed the scope and availability of molecular diagnostic tests in the National Health Service. While the technology has provided access to novel wet laboratory workflows and bioinformatic tools, understanding their limitations remains essential to being able to provide safe and effective services for patients.

This body of work, which was undertaken as part of ongoing research and development priorities within a busy Translational Genomics Laboratory, reports progress towards developing, evaluating, and assessing techniques that have the potential for clinical utility. Investigations were initially focussed on development of a "single cell" analysis workflow for the identification and characterisation of structural (copy number) changes that could support and ultimately replace array-based embryo screening assay. Subsequent work sought to investigate a novel target enrichment reagent to interrogate regions of the genome that are refractory to analysis by short-read sequencing. The *PMS2* locus, which is associated with hereditary predisposition to colorectal cancer, was selected for this purpose. The research, which used hybridisation capture of "linked-read" whole genome sequencing libraries, proved successful, but progress was halted due to the proprietary reagents being withdrawn from sale following a legal dispute between vendors. As time progressed and so-called "third generation" long-read sequencers were developed, the third and final area of research sought to assess the utility of nanopore sequencing for the diagnosis of rare disease patients that had been sequenced using the short-read next-generation sequencing workflow.

1.2 Early approach to genetic testing

In 1958, Francis Crick published his seminal paper "On Protein Synthesis" (Crick, 1958) which proposed the Central Dogma of Molecular Biology, describing the flow of genetic information within a biological system. This framework outlines the process by which DNA is transcribed into RNA and then translated into proteins, highlighting the relationship between genes and the traits they encode. These insights profoundly influenced the field of genetics, shaping our understanding of how genetic information is expressed and regulated in living organisms. It was around this time that the first genetic tests began to emerge, prompted by advancements in our understanding of DNA, its packaging into chromosomes and the resulting genetic basis of disease.

In the early-mid 20th century, improvements to cell culturing techniques, in particular the inclusion of colchicine to induce cell cycle arrest in metaphase, led to the development of karyotyping, with confirmation in 1956 that euploid germline cells in humans possess 46 chromosomes (Tjio & Levan, 1956). This was followed in 1958 by the discovery of the first human aneuploidy, Down syndrome (Lejeune *et al.,* 1959). The characterisation of sex chromosome abnormalities, including Klinefelter's (Jacobs & Strong, 1959) and Turner syndrome (Ford *et al.,* 1959) were described soon after. The introduction of chromosome banding a decade later heralded the beginning of diagnostic cytogenetics, with the ability to differentiate chromosomes and detect smaller gains and losses now possible (Ferguson-Smith, 2015).

Due to the inability to directly interrogate targeted DNA sequence, early diagnosis of single-gene disorders focused on pedigree analysis. In 1977, Fred Sanger developed the chain termination sequencing technique, which would ultimately become known as "Sanger sequencing" (Sanger, Nicklen and Coulson, 1977). This used dideoxy nucleotides, chemical equivalents of standard deoxynucleotides without the three-prime hydroxyl group that allows incorporation of the subsequent nucleotide. These labelled ddNTPs were added to a mixture of regular dNTPs, allowing the synthesis of DNA strands of varying lengths to the point a ddNTP is incorporated. Size separation by polyacrylamide gel electrophoresis, and later capillary electrophoresis instruments, as well as the invention of the polymerase chain reaction in 1986 (Mullis *et al.,* 1986), gradually improved sequencing throughput.

Within the first two decades of the 21st century, short-read next generation and long-read third generation sequencing instruments have been developed by numerous manufacturers. These devices can perform whole genome sequencing in a matter of days and have expanded the scope of genetic testing across clinical specialities and spectrums of development (from reproductive and pre-natal screening to paediatric and adult medicine).

1.3 Sequencing technologies

Since the development of automated capillary electrophoresis or "Sanger" sequencing numerous high throughput sequencing platforms have been developed, these are categorised as either short-read next-generation or long-read third generation instruments, depending on the length of molecule that can be sequenced. To ensure samples are ready for sequencing it is necessary to manipulate DNA fragments so that they contain sequencer-compatible adaptors; this process is ubiquitously termed "library preparation". Typically, it involves either an adaptor molecule being ligated to the ends of the fragments, or a transposase being used to incorporate the adaptor into a double stranded locus. And depending on the selected instrument it is necessary for this to happen before (hybridisation capture), during (smMIPS), or after (LR-PCR) target enrichment.

1.3.1 Short-read next generation sequencing

Numerous next-generation sequencing (NGS) platforms have been developed over the past two decades. While these include instruments such as the 454 (Roche) and SOLiD (ABI), the most dominant and ubiquitous series of platforms use the sequencing-by-synthesis chemistry commercialised by Illumina (Bentley *et al.,* 2008).

Illumina-compatible adaptors are added to DNA fragments (or inserts) and the pooled libraries are sequenced on an instrument-specific flow cell. This is a plastic, or glass, consumable that contains between 1, 2 or 4 lanes coated with a lawn of oligonucleotides. While Illumina instruments differ in the yield or length of sequence data that can be generated from the flow cell, they all perform sequencing-by-synthesis reactions. Once the complementary sequencing adaptors have bound to the flow cell the target is bridge-amplified to create a cluster and after PCR-extension one end is dissociated. A sequencing primer binds to the exposed adaptor then iterative rounds of fluorescently labelled dNTPs are washed across the flow cell, extending the primer sequence one complementary nucleotide at a time. In between each additional extension reaction, the flow cell is imaged; the number of consecutive cycles defines the length of the sequence read. For short-read Illumina sequencing this is typically anything from 50 to 150 bp. A further characteristic of short-read Illumina sequencing is the ability to subsequently read the same molecule from the opposite end, generating a "reverse" read. Depending on the size of the insert the forward and reverse reads may, or may not, overlap. Either 1 or 2 index reads (depending on library type) are performed to allow the sequence read to be attributed to a source sample.

The latest range of models currently includes the low-throughput MiSeq i100, medium-throughput NextSeq1000/2000 and high-throughput NovaSeq X/Plus instruments. Improvements including a patterned flow cell, room temperature reagent shipment and chemistry enhancements have served to progressively improve the speed, quality, yield, and end-user experience. Such developments are iteratively incorporated into updated releases of each model over time. This has significantly reduced the cost-per gigabase of sequence making whole genome sequencing a possible, but data intensive endeavour.

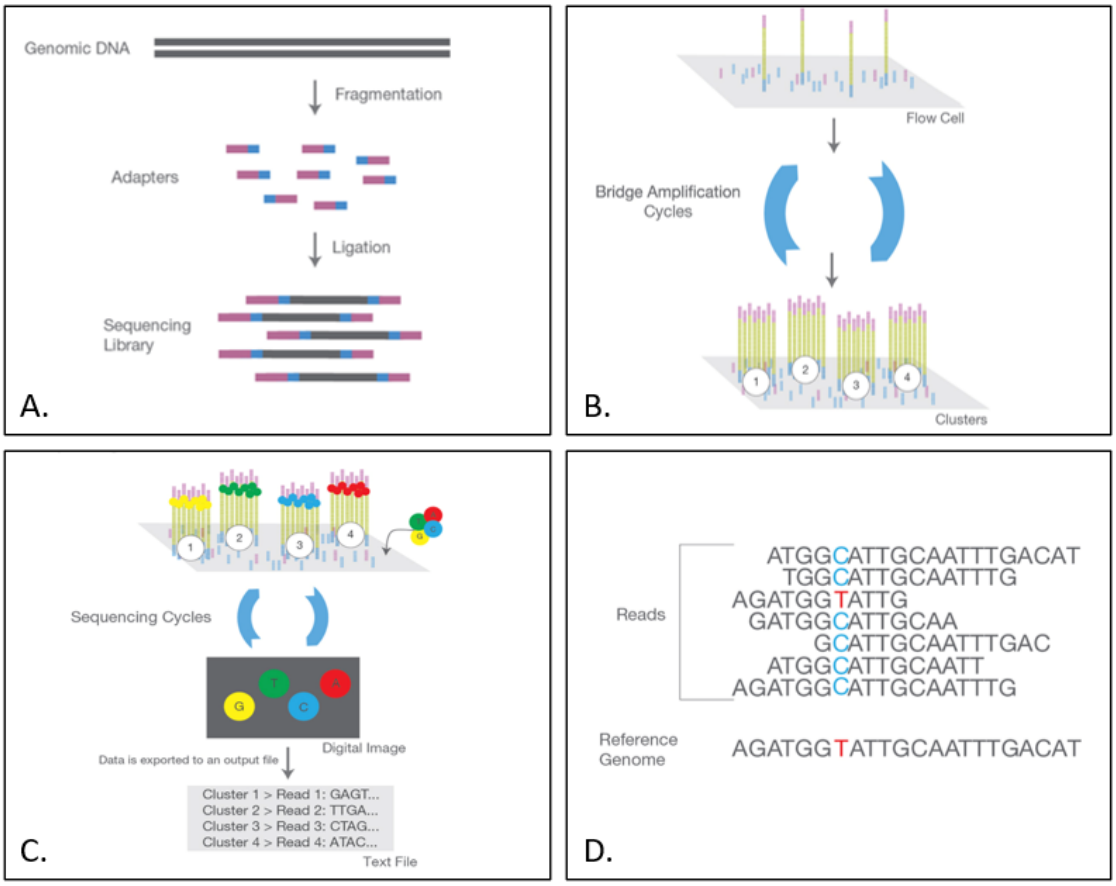

Figure 1.1: The process of short-read Illumina sequencing. (A) Genomic DNA is fragmented to approximately 200-300bp in length and adapters are ligated onto each end to generate a library of sequencing compatible fragments. **(B)** When the library is added to the flow cell, the sequencing adapter hybridises to the complementary sequence on the inside surface of the flow cell lane. Library fragments are bridge amplified to generate a clonal cluster of the originating molecule. A sequencing "read" is generated by synthesis of a complementary fragment along the library template. **(C)** A single fluorescent nucleotide is added per cycle, imaged, and the fluorescent signal converted into a base call file. **(D)** The raw base call data is converted into file formats compatible with downstream analysis software, where the sequencing reads are aligned to the reference genome for comparison. Image courtesy of Illumina, Inc.

1.3.2 Target enrichment

To enable the capacity of short-read sequencing instruments to be used in a timely and cost-effective manner a subset of loci is typically targeted for sequencing. Numerous methods for target enrichment have been developed, each with specific strengths and weaknesses. Regardless of the target enrichment approach used, sample indexing, which allows multiple patients to be pooled and sequenced concurrently is a universal feature common to all strategies.

1.3.2.1 Long-range PCR

One of the simplest enrichment strategies is long-range PCR across the target locus followed by sequencing of the amplicons; this approach has been used to effectively screen numerous disease-associated genes including those for retinitis pigmentosa (de Sousa Dias *et al.,* 2013) and hereditary cancer (Carr *et al.,* 2013). The use of optimised DNA polymerases enables up to 30-kb of sequence to be amplified in a single long-range PCR (LR-PCR) reaction (Jia *et al.,* 2014). When the genomic region of interest is small, amplification products can be generated in a quick and cost-effective manner, with a well-designed strategy this can be performed efficiently in 96-well plates. Multiple fragments are typically combined in equimolar concentrations for sequencing. Nevertheless, there remains a finite limit to the number of amplicons that can be efficiently processed, preventing the approach from being indefinitely scalable. Furthermore, the high molarity of amplification products has prompted concerns that repeated use of a limited number of primer pairs could lead to low-level contamination within the laboratory.

An additional limitation of LR-PCR enrichment which reduces test sensitivity, is the unidentified presence of a primer-site variant that results in non-amplification of one allele. The absence of heterozygous genotypes across assayed LR-PCR fragments (which supports the amplification of two alleles), makes it difficult for high-throughput clinical laboratories to quantitate this risk of misdiagnosis due to allele dropout, instead pro-active steps typically taken to minimise this risk including the use of variant databases to ensure there are no polymorphic variants in the primer sites. Additional considerations that can affect the robustness, reliability and ultimately test sensitivity for LR-PCR enriched amplification products include the presence of repeat sequences, the formation of secondary structure (Lam & Mak, 2013) such as hairpin loops or the GC content or the target region.

1.3.2.2 Hybridisation capture

To perform enrichment for a larger proportion of the genome than is possible using LR-PCR amplification hybridisation capture methods were developed. For this strategy a custom reagent of long (~100 bp) RNA or DNA probes are synthesised that are complementary to the regions of interest. After genomic DNA has been sheared and had sequencer-compatible adaptor sequences ligated to their termini, the probes bind to complementary sequences. The probes themselves are biotin labelled allowing the DNA/probe heteroduplex to be bound to streptavidin beads; the complex is captured using a magnet and enriched by PCR, generating fragments that are suitable for sequencing.

The first hybridisation capture studies demonstrated scalability of the technique which enabled whole exome sequencing of all target exons. Numerous manufacturers have commercialised the workflow with their own off-the-shelf probe sets for different types of exome reagent and bespoke panels. For example, some reagents focus on all disease-causing genes (often termed the clinical exome) whereas other reagents focus on specific clinical specialties (*e.g.* a regent for cardiomyopathies genes or hereditary cancer). Whole exome sequencing (with probes targeting all the coding sequences of the genome) enables scrutiny of recurring variants that align to patient phenotype in novel loci, and thereby this technology has been responsible for the majority of new disease gene associations in recent years.

Common to all reagents is the impact of emerging knowledge; new disease gene associations, candidate loci, or interesting variants located in previously uncovered regions (such as disease-causing variants causing cryptic splice sites or impacting regulatory sequences) are frequently reported, rendering existing probe designs out-dated. As is the case with all target enrichment regions, laboratories (especially those performing diagnostic screening) have to regularly update their designs and validate the performance of new reagents.

Vendors provide software that facilitates probe design. With such widespread adoption of hybridisation capture workflows, the predicted performance of designed probes is now robust. Genomic regions with high GC content (frequently including a gene's first exon) have been characteristically hard to sequence due to an inability to denature and hybridise probes using standard conditions (Clark *et al.,* 2011). With a required minimum sequencing depth of at least coverage for germline variant calling, the empirical data that has been generated using these reagents has provided information about which regions of the genome would benefit from additional probes to boost sequencing performance. These data have also improved the uniformity of coverage such that the eventual number of captured sequence reads is less variable across the reagent (regions of high sequence coverage are reduced and regions of low sequence coverage are increased); this has the effect of requiring less sequence data per sample lowering the associated sequencing cost (either by being able to pool more samples together or using a smaller cartridge).

By contrast to LR-PCR enrichment, it is possible to identify copy number variants in samples prepared by hybridisation capture. Specialist software, such as ExomeDepth, allows the number of reads occurring in a given genomic region to be counted and compared to a panel of so-called normal controls (cases that are presumed to not harbour a copy number variant in the interrogated gene). Comparative read depth analysis has increased the mutation spectrum for many diseases and in a study performed in our laboratory we were able to identify a series of multi-exon deletions in patients that had been referred for analysis of Joubert or Meckel Gruber syndrome (Watson *et al.,* 2016a). These patients had single *bona fide* heterozygous variants identified in disease-associated autosomal recessive genes that matched their clinical phenotype. Prior to analysis by comparative read depth analysis a second pathogenic allele had not been identified so a molecular diagnosis could not be confirmed. The utility of comparative read depth analysis is now widely acknowledged and integrated into data processing workflows that analyse hybridisation capture datasets.

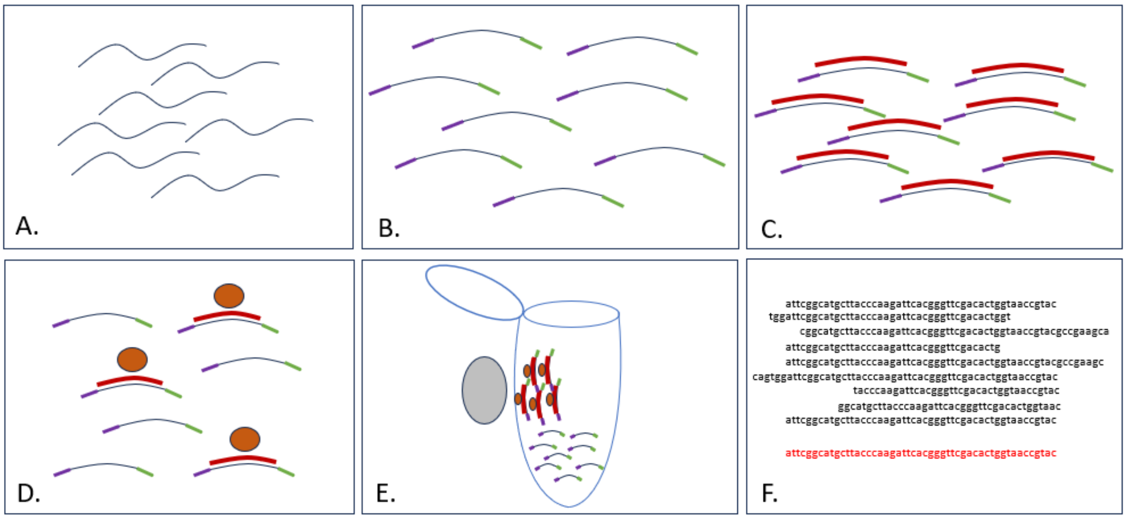
A.
B.
C.
D.
E.
attcggcatgcttacccaagattcacgggttcgacactggtaaccgtac
tggattcggcatgcttacccaagattcacgggttcgacactggt
cggcatgcttacccaagattcacgggttcgacactggtaaccgtacgccgaagca
attcggcatgcttacccaagattcacgggttcgacactg
attcggcatgcttacccaagattcacgggttcgacactggtaaccgtacgccgaagc
cagtggattcggcatgcttacccaagattcacgggttcgacactggtaaccgtac
tacccaagattcacgggttcgacactggtaaccgtac
ggcatgcttacccaagattcacgggttcgacactggtaac
attcggcatgcttacccaagattcacgggttcgacactggtaaccgtac
attcggcatgcttacccaagattcacgggttcgacactggtaaccgtac
F.

Figure 1.2: Hybridisation capture target enrichment workflow. (A) DNA samples are first sheared. **(B)** Sequencer compatible adapters are ligated to the ends of fragments to create a library. **(C)** Biotinylated DNA or RNA baits (oligos which complement the regions of interest) are hybridised to the library. **(D)** Streptavidin beads bind to the baits. **(E)** A magnet is used to separate the captured fragments. **(F)** The captured library is enriched by PCR and taken forward for sequencing, with the majority of reads generated aligning to the target loci.

1.3.2.3 smMIP enrichment

For screening large cohorts of samples for a modest number of loci (10's of genes) the single molecule molecular inversion probe (smMIP) workflow has proved robust and popular. Here, individual single stranded DNA probes are designed to target each region of interest. The probes comprise a common 30 nucleotide linker backbone, a 5-nucleotide sample index, two degenerate 5 nucleotide sequences and two target specific 16-24 nucleotide extension and ligation arms.

The extension and ligation arms hybridise complementary DNA and flank a 225 bp region of interest. The extension arm acts as a primer to allow in-filling across the region of interest, and a ligase enables the formation of a circularised probe. Many custom-designed probes are circularised concurrently in a multiplex reaction. Following this process, the remaining linearised genomic DNA, and smMIPs that did not hybridise are removed by exonuclease treatment. The circularised probes are then PCR-enriched using universal primers that bind to the backbone, resulting in patient-specific linearised products; these are quantified and pooled in equimolar concentrations for sequencing.

Molecular inversion probe protocols have been iteratively refined. The current methodology, which includes a 5-bp degenerate sequence allows discrimination between probes that are PCR-replicates of each other and probes that have bound to unique DNA fragments. This presents the possibility of being able to use smMIPS to detect variants that are somatic in origin, and which are identified by their low variant allele fraction (i.e. not the 50:50 ratio that is expected for a germline DNA sequence variants) (Hiatt *et al.,* 2013). Early uses of inversion probe technology included the genotyping of single nucleotide polymorphisms (Hardenbol *et al.,* 2003) before it was reported that the approach could be scaled to capture and amplify approximately 10,000 human exons in a single multiplex reaction (Porreca *et al.,* 2007). Local experience of smMIPs has focussed on analysis of *ABCA4*, a 128-kb gene that is the single most common cause of inherited retinal dystrophy (Mc Clinton *et al.,* 2023). Our multi-centre international study was made possible due to the low ongoing cost of deploying the smMIPs workflow; once the probes have been synthesised many thousands of samples can be analysed from the source reagent (Hitti-Malin *et al.,* 2022).

Despite the long-term low-cost of using an smMIPs pool the enrichment workflow has a number of limitations. This includes the manual rebalancing and difficulty of introducing new probes into an established working reagent. Also, similar to LR-PCR and hybridisation capture enrichment approaches, it is not possible to establish the phase of identified variants (determine whether they are arranged in *cis* or *trans*) unless they are positioned sufficiently close to each other to have been captured in the same sequencing read or cascade testing can be performed in other family members.

1.3.2.4 CRISPR enrichment

In contrast to the target enrichment methods described thus far, Cas9 cleavage offers the possibility of enriching genomic DNA using a PCR-free approach. This has the benefit of being able to both generate target molecules that are longer than can be generated from a standard or long-range PCR reaction and have their methylation status preserved. Use of these enriched fragments on instruments that support the sequencing of long DNA molecules will likely be of future interest.

In practice, the workflow relies on components of the bacterial clustered regularly interspersed short palindromic repeats (CRISPR) system. By designing a unique crRNA and combining it with a generic tracrRNA and CRISPR-associated protein (such as *Streptococcus pyogenes* Cas9) a ribonucleoprotein (RNP) complex can be generated. When incubated with genomic DNA the RNP will generate double strand breaks, the ends of which can be prepared for sequencing.

Genomic regions that contain tandemly repeated sequences are attractive targets for enrichment using a CRISPR workflow. At these loci the size of a repeat-containing allele can exceed the capability of LR-PCR. Amplification reactions also typically result in preferential amplification of the smaller allele resulting in allele dropout. While southern blotting has historically been used to assess these loci it is now known that repetitive tracts are frequently imperfect, containing interruptions that may influence the stability of the repeat (its ability to contract and expand over time). And with recent advances in the culturing and differentiation of primary cells it is now possible to assess the stability of repetitive tracts from a variety of tissues. In a study of Fuchs endothelial corneal dystrophy (an age-related cause of vision loss) focussed on repeat stability at the *TCF4* locus (the most common repeat expansion-mediated disease in humans) a CRISPR-guided workflow aided the identification of alleles containing more than 1,500 repeats (Hafford-Tear *et al.,* 2019).

Molecular diagnostic use of CRISPR enrichment workflows have also been demonstrated for the single-nucleotide characterisation of structural variants. In a study performed by our laboratory we sited guide RNAs within an apparently duplicated sequence and were able to generate sufficiently long DNA sequences to resolve the variant breakpoint (Watson, Crinnion, *et al.,* 2020). This application had utility in determining whether a copy number variant was classed as a duplication (which has one identifiable breakpoint) or insertion elsewhere in the genome (which has two identifiable breakpoints). The importance of characterising the genomic architecture at the source locus is important as this can lower the risk of pathogenicity of an identified variant if it is inserted into a site elsewhere in the genome with no known disease-associated features.

In our experience establishing the efficiency of probe cleavage, for ad-hoc assays, can be challenging. While optimisation experiments can be performed on PCR amplification products (or commercially synthesised double stranded DNA molecules) that contain the guide cRNA target, they correlate poorly with use of the same RNP in combination with genomic DNA samples that comprise many more species. Consequently, to ensure robust cleavage, it is often necessary to generate RNPs using a collection of guide cRNAs. A seemingly better approach is to empirically test, and then iteratively design, guide cRNAs at designated target loci. Inevitably these panels are expensive and time-consuming to optimise. Nevertheless, for neurodegenerative disease-associated repeat loci, PureTarget^{TM}, which was developed and commercialised by PacBio, was recently launched: Up to 48 samples can be assayed concurrently at 20 expansion loci using as little as 1 µg of DNA (PacBio, 2025).

1.3.2.5 PCR-free library preparation

For many molecular diagnostic screening workflows, the sequencing cost now represents a small proportion of the overall test cost (when the technical and scientific staff time relating to the production and interpretation of test results have been considered). This access to “cheap” sequence data has meant target intervals are frequently over-sequenced with more data being produced than is necessary. Nevertheless, for WGS assays the large size of the human genome has required ongoing optimisation of library preparation and instrument performance to ensure studies remain within budget. One such advance is the move to PCR-free libraries; while amplification bias is inherent to short read sequencing technology, PCR-induced bias is eliminated from library preparation, and more even coverage can be obtained across high-GC or -AT regions. This is often achieved using tagmentation workflows whereby a transposase can integrate the adaptor sequences directly into the genomic DNA sample in a single step, reducing library preparation to a 90-minute process. Furthermore, because tagmentation occurs on an immobilized bead the need to perform pre- and post-library preparation quantification is eliminated. While WGS is being widely adopted, including by the NHS, some disadvantages exist. The most obvious of these which affects the sensitivity of variant detection, is the reduction in absolute read coverage across a particular locus. Additionally, the use of DNA sources such as saliva samples can, in the absence of a target enrichment method, result in a dataset that has extensive bacterial contamination.

1.3.2.6 Linked-read library preparation.

While third-generation sequencers are making it possible to sequence long (>150 bp) DNA fragments several vendors have commercialised technology that allows synthetic long-reads to be generated utilising existing short-read infrastructure. Conceptually these methods tag DNA molecules using a fragment-specific barcode prior to them being sheared into smaller fragments for library preparation. Following short read sequencing the fragment-specific barcodes can be re-constituted to assemble multi-kilobase “linked” reads.

By using linked-read sequencing it is possible to arrange identified variants into haplotypes. For autosomal recessive disorders this provides an opportunity to identify biallelic variants occurring in *trans* (on different parental haplotypes) that are more likely to be disease causing. Furthermore, linked reads can aid the mapping of sequence reads that would otherwise be ambiguously aligned. These sites are frequently within repetitive regions and from a molecular diagnostic perspective are often the sites of structural variant breakpoints.

One of the early leaders of linked-read mapping technology was 10X Genomics (https://www.10Xgenomics.com). The 10X linked-reads workflow uses the Chromium controller to facilitate bead-based barcoding of fragments inside a gel bead-in emulsion

(GEM) which partitions the genomic DNA. The DNA fragments that share the same barcode sequence are grouped together as they originate from the same single DNA molecule. In order to re-constitute the source DNA molecules, the Longranger bioinformatics data processing pipeline was developed by 10X Genomics. To appropriately partition the genome only nanogram quantities of DNA are required to make the library.

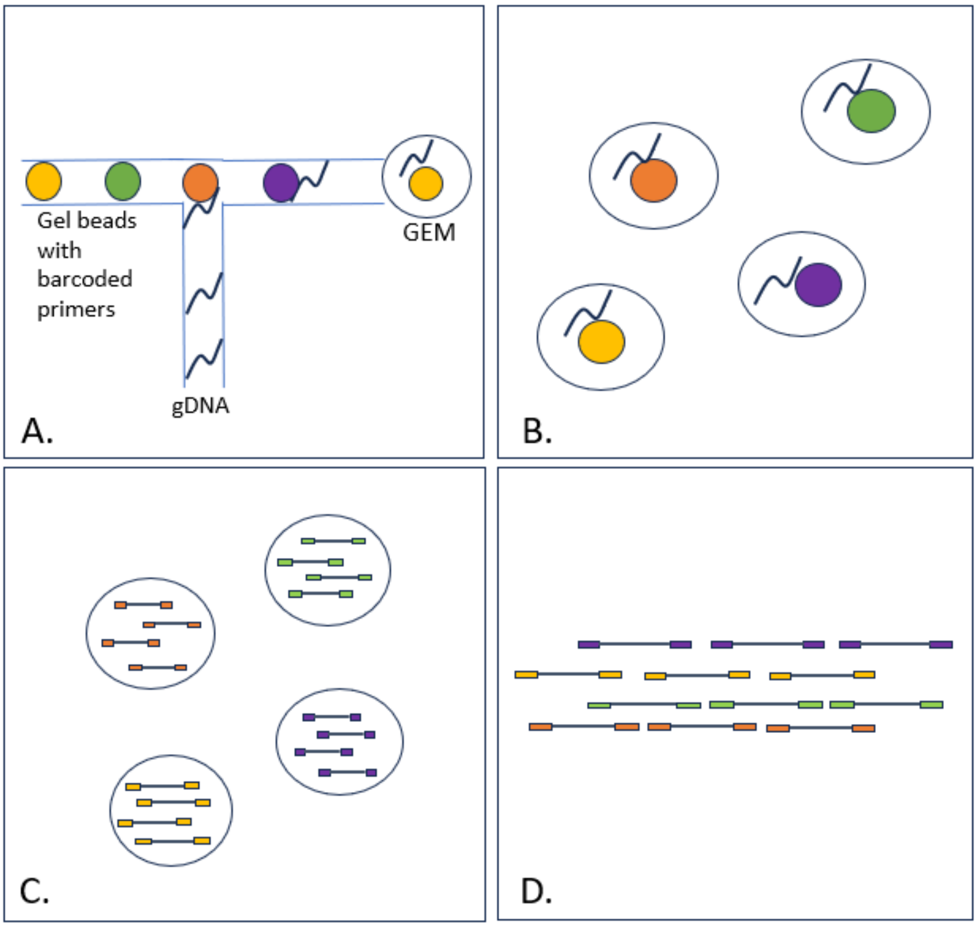

Figure 1.3: Schematic overview of the 10X Genomics linked-read sample preparation workflow. (A) Inside the Chromium cartridge, HMW DNA is forced through narrow channels, where DNA molecules are partitioned into oil droplets with the gel beads. **(B)** Each GEM (Gel bead-in emulsion) contains unique barcoded primers. **(C)** The beads are dissolved to release the primers, and the DNA is amplified into barcoded fragments. **(D)** Following library preparation and short-read sequencing, the unique barcodes are used to link the sequencing reads and re-construct the source DNA molecule. Image adapted from (10X Genomics, 2019)

The first use of linked read technology demonstrated haplotype block phasing for a nuclear trio with genotypes that were concordant with expected inheritance patterns (Zheng *et al.,* 2016). Furthermore, exome sequencing was applied to a linked-read whole genome library allowing the structure of an *EML4/ALK* fusion to be resolved in the NCI-H2228 cancer cell line (Zheng *et al.,* 2016).

More recently, constellation mapped read technology promises to improve the mapping of challenging genomic regions through the creation of a synthetic long read.

The technology, which has been developed by Illumina, uses a library preparation process that occurs on the surface of a short-read sequencing flow cell. Transposomes are first bound to the flow-cell surface before double stranded DNA is pushed through the flow cell and tagmented. This results in DNA being bound to the nanowells on the flow cell. Standard short-read sequencing is then performed with the sequence data from adjacent wells linked to allow a synthetic long-read to be constructed. Although the constellation workflow is not due to be released until 2026, its compatibility with existing NovaSeq X systems suggests it will be an exciting future product (Illumina, 2024). Exemplar datasets suggest N50 phase blocks can reach ~715 kb with standard DNA extractions and ~5.7 Mb with high molecular weight extractions. Furthermore ~98% of heterozygous single nucleotide variants could be phased by both standard and high molecular weight extractions. If the purported capabilities of this technology are proven when implemented in the wider sequencing community, this could allow Illumina to compete with the more established long read sequencing technologies in the clinical arena, due to the majority of diagnostic labs already having this infrastructure in place.

1.3.3 Long-read single molecule real-time (SMRT) sequencing

Although synthetic long-read library preparation, in combination with short-read sequencing offers the potential to generate a phased genome and better detect and characterise variants in challenging regions of the genome, a more conceptually satisfying approach is to perform native sequencing of a much longer DNA molecule. Several long-read sequencers have now been developed including “single molecule real-time sequencing” commercialised by PacBio. For many years base-calling accuracy was far worse than that which could be achieved for Illumina sequencing, however so called “HiFi reads” are now able to achieve 99.9% single read accuracy.

SMRT sequencing is performed using a SMRT Cell chip which contains millions of small wells within which the reactions occur. In a similar manner to short-read workflows, libraries are prepared by ligating adaptors to the termini of DNA fragments which are typically ~15-20 kb in length. A notable difference is that the PacBio adaptors are hairpin in topology and create a circular DNA molecule, known as a SMRTbell library. Sequencing primers bind to the adaptors, and as each new base is incorporated using a DNA polymerase and complementary base pairing, fluorescent light signals are emitted. As the molecule is circular, the DNA polymerase can repeatedly pass around the fragment. For each fragment a dataset is created that consists of adaptor sequence followed by sense strand sequence, adaptor sequence and finally antisense strand sequence. When the DNA polymerase returns to the primer binding site the process is repeated. After the run has completed, adaptor sequences are trimmed from the "polymerase read" to generate a series of "subreads". The subreads are combined to create a consensus, highly accurate, HiFi read, leaving only stochastic sequence errors which are usually attributed to issues of fluorescent signal detection or polymerase lag.

There are two available PacBio platforms, the high-throughput REVIO instrument that can run up to 4 SMRT Cell consumables concurrently, yielding 100-120 Gb per Cell in 24 hours. HiFi reads allow whole genome sequencing data to be manufactured at a lower per-base read depth (typically 20×) compared to short-read sequencing (typically 30×), meaning less sequence data is required to reach equivalent variant calling sensitivity. On this basis, two WGS samples are run per SMRT Cell. By contrast the lower throughput VEGA instrument is capable of generating 60 Gb per SMRT cell (1 WGS sample) (PacBio, 2025).

To generate long-read sequences, third generation workflows have typically focussed on the sequencing of genomic DNA, while this allows the native detection of methylation status (without sodium bisulphite treatment) the required mass of input DNA has often been significant. For many protocols at least 1-2 µg of DNA is required for library preparation, but due to the need to perform pre-treatment reactions, using short-read eliminator (which removes fragments less than ~10 kb), a much greater starting mass is often required. An incremental improvement to SMRT sequencing includes a soon-to-be-released SPRQ chemistry which will lower the DNA requirement for library preparation to as little as 500 ng.

To address the data storage demands posed by high-throughput sequencing PacBio has developed a highly efficient approach to data storage. The storage requirement is 0.5 bytes/base of raw data (REVIO BAM format), this in contrast to competitor long-read platforms which require 10.0 bytes/base of raw data (FAST5/POD5 format) (PacBio, 2025). With many sequencing companies now focussed on the clinical adoption of their platforms, reductions to data storage requirements will likely be a platform characteristic that is welcomed by hospital IT departments.

1.3.4 Long read nanopore sequencing

In contrast to SMRT Cell sequencing, Oxford Nanopore Technologies have developed a range of long-read sequencers that detect DNA molecules passing through membrane-embedded nanopores. Akin to other next and third generation workflows, libraries are loaded onto a disposable consumable for sequencing, which takes place using either a low throughput MinION, medium throughput GridION or high throughput PromethION.

Instruments detect the analyte as a disruption to the hydrogen ion gradient that is established across the membrane. The combination of nucleotides in the nanopore "reader" generates a characteristic "squiggle" plot that can be base called into a conventional sequence read. By contrast to the PacBio platform which determines methylation status following an analysis of polymerase kinetics, nanopore sequencing identifies methylation (N6-methyladenine, 5-methylcytosine and 5-hydroxymethylcytosine) directly from its representation on the squiggle plot (Oxford Nanopore Technologies, 2024). This unique approach has allowed the development of protocols that can natively detect RNA and associated base modifications (including N^6-methyladenosine and pseudouridine (Hassan *et al.,* 2022).

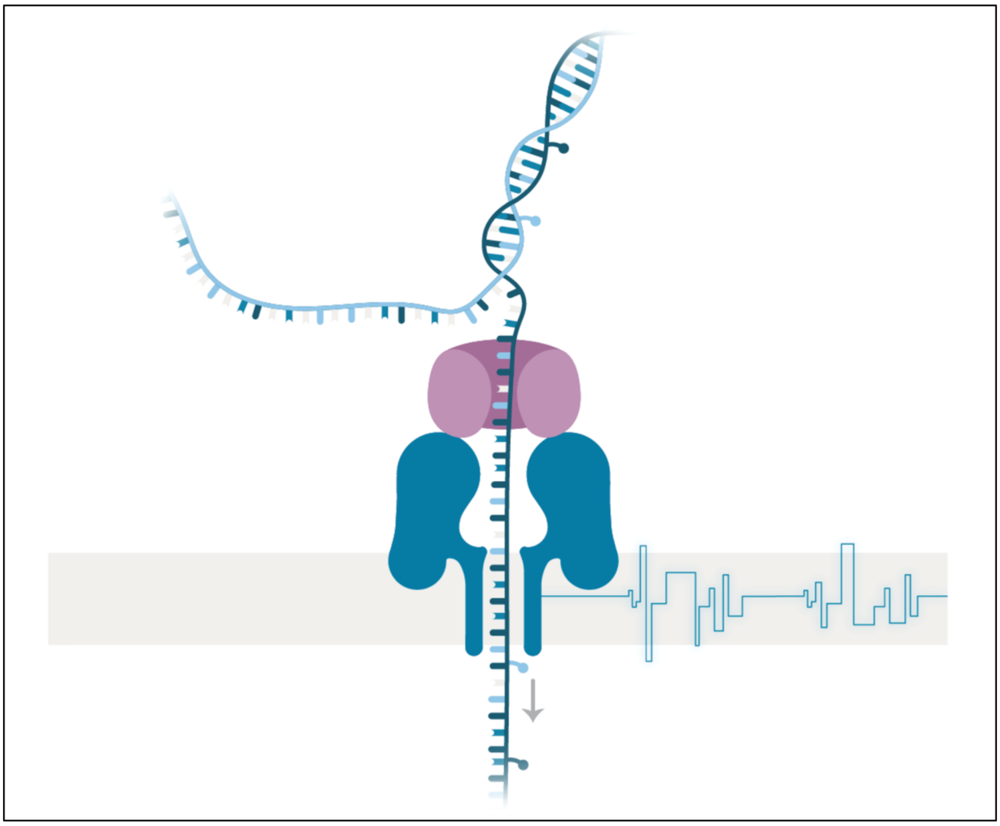

Figure 1.4: Diagram of Oxford Nanopore Technologies sequencing. Native DNA strands pass through the nanopore, disrupting the current to generate the characteristic squiggle plot from which nucleotide sequence is inferred. Image courtesy of Oxford Nanopore Technologies.

Nanopore platforms do not limit the maximum length of sequence reads. This distinguishing characteristic has led to the development of ultra-long sequencing protocols and reports of megabase-length sequence reads (Jain *et al.,* 2018). The utility of ultralong reads has been demonstrated by the Telomere-to-Telomere consortium, a community-based effort to generate the first complete genome assembly. Sequencing the human CHM13hTERT cell line, which has closed many of historical gaps, has created a new reference standard, T2T CHM sequence. Much work has been undertaken to analyse telomere and centromere sequences, regions of the genome that are comprised of many tandemly repeated segments.

Deploying ultralong read protocols in a clinical setting remains challenging; the handling of high molecular weight DNA requires specialist extraction protocols which cannot currently be accommodated by routine DNA extraction protocols implemented in pathology laboratories. In an attempt to overcome this challenge, the SageHLS^{TM} instrument (Sage Science) was developed to enable Cas9 enrichment of large genomic fragments directly from white blood cells. Cells are lysed under electrophoretic conditions and pulse-field gel electrophoresis in combination with Cas9 cleavage allows subsequent electro-elution from a SageHLS cassette. While the workflow has yet to be implemented into routine clinical use there are reports of it being successfully used to characterise segmental duplication-mediated deletions at the 22q11 locus (Zhou *et al.,* 2024).

For many laboratories nanopore sequencing is the most cost-effective technology for accessing long-read sequencing. The cheapest low-throughput consumable a “Flongle” flow cell generates up to 2 gigabases of sequence data for a flow cell cost that is as low as £75.00. This has facilitated significant uptake in the microbial sequencing community, where information about antibiotic resistance and the entire complement of bacterial species in a given sample is helping amend clinical pathways. Importantly, although data heavy, the nanopore file format can be read in real-time providing immediate feedback to analysts. In addition to microbiology applications, this real-time sequencing analysis is being studied as part of intra-operative investigations of brain cancer. Here the idea is to perform a molecular classification of the tumour using an adaptive sampling enrichment; DNA fragments that enter the pore from regions of the genome that are not selected are ejected by reversing the current. Rapid intra-operative sequencing presents the possibility of using a single workflow to obtain all molecular information in real-time and from a single specimen, potentially eliminating the need to perform immunohistochemistry or EPIC arrays to obtain the methylation status.

Prior to the availability and competitive pricing of the high-throughput PromethION instrument, adaptive sampling was demonstrated to be effective for selective sequencing of rare disease referrals. Panels of genes could be selectively screened to identify pathogenic alleles that were refractory to detection by short read technologies. While adaptive sampling workflows are now supported on the GridION the hardware requirements remain significant – sequence reads need to be aligned to reference genome in real-time while the fragment is passing through the pore at approximately 450 bases per second. While all that is necessary to initiate the selective enrichment (or depletion) of DNA fragments is a three-column bed formatted file comprising the chromosome, start and stop coordinates on the reference genome, the varying performance between different bed files suggests that whole genome sequencing on a high-throughput instrument may be a more effective approach to increasing diagnostic yield in rare disease patients.

1.4 The landscape of genetic testing in England

Over the course of this work, the geographical distribution and approach to genetic testing in the UK National Health Service has been transformed. Now, instead of services being contracted through a network of approximately 20 regional laboratories, work is commissioned through 7 Genomics Laboratory Hubs (GLH) serving populations that are comparable in size (Figure 1.5). Each GLH has the facilities and skills to offer a wide range of tests, such as karyotyping, aCGH, microsatellite genotyping, sizing of repeat expansion, southern blotting, and DNA sequencing. In addition to these laboratories diagnostic whole genome sequencing data is produced in the UKAS accredited Illumina Laboratory Services (ILS) Laboratory, on the Hinxton campus. While the manufacturing of these data is beyond the scope of normal NHS activity, the identified variants are interpreted in a diagnostic environment that returns findings to clinicians and their patients. Nevertheless, private sector involvement in molecular genetic diagnosis remains scarce in the UK, notable exceptions being in reproductive health for previously discussed PGS via aCGH and non-invasive prenatal testing for fetal trisomy detection.

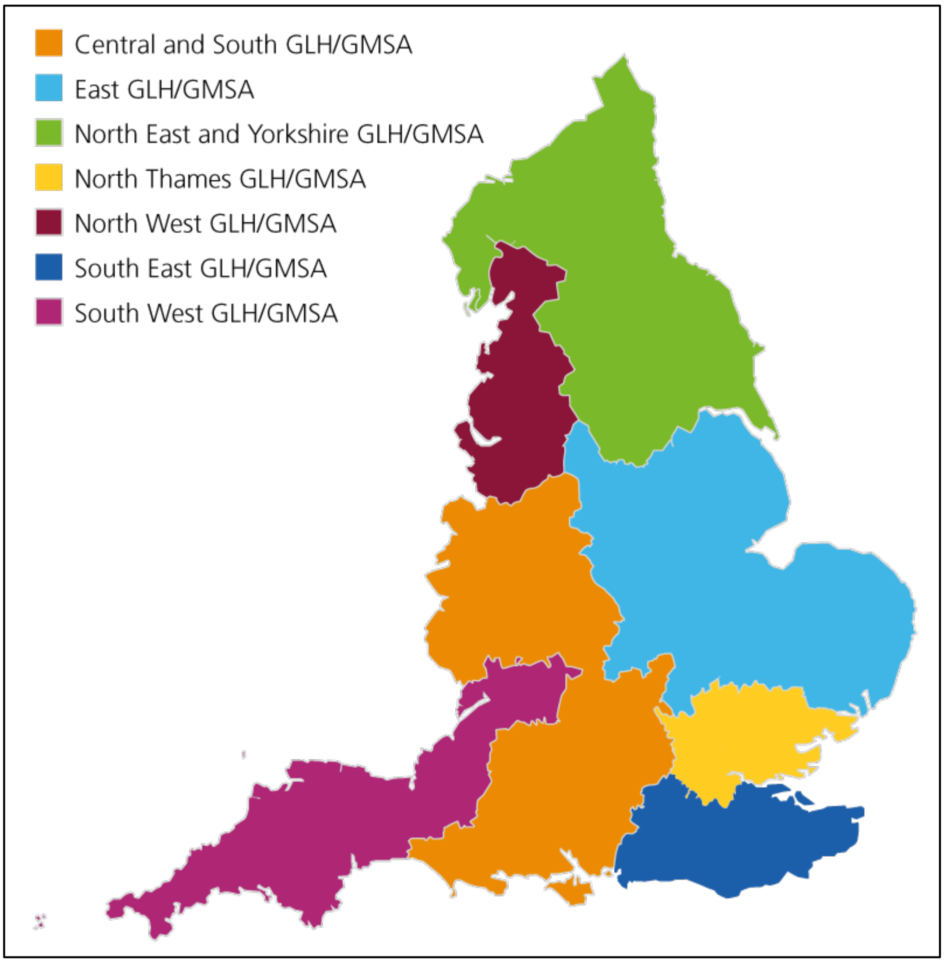

Figure 1.5: The distribution of the seven Genomic Laboratory Hubs in England. The regions work collaboratively with each other and are named as per the key. The clinical care for patients receiving genomic testing is provided by the Genomic Medicine Service Alliance (GMSA) that is aligned to the corresponding GLH. Image reproduced from: https://www.genomicseducation.hee.nhs.uk/genotes/knowledge-hub/genomic-laboratory-hubs/

For diagnostic tests that are performed by short-read sequencing, (and for whom WGS is not performed), at least one or more GLHs perform the wet-laboratory preparation and sequencing of the patient samples. Frequently this involves DNA samples being re-distributed to an appropriate GLH laboratory for testing; in some scenarios (particular for the identification of de novo variants in neurodevelopmental disease) trio testing is performed, this involves sequencing the proband and their parents.

1.5 Genomic analysis in reproductive medicine

Approximately 30% of clinically recognized pregnancies progress to liveborn babies; when pregnancy loss occurs, this is most frequently due to chromosomal abnormality in the embryo (Macklon *et al.,* 2002) When conception occurs following the use of assisted reproductive technology (ART), there is an opportunity to screen the pre-implantation embryo prior to transfer. Several criteria have been proposed with the primary one being embryo morphology (Sigalos, Triantafyllidou and Vlahos, 2016); the aim is to improve conception rates and pregnancy outcomes. Since the mid-1980s, research has been ongoing to determine how genetic analyses can be best integrated into the screening process (Delhanty, 1994).

1.5.1 Preimplantation genetic diagnosis and screening

The earliest use of genetic testing in the field of assisted reproduction, was sex-selection of embryos with the aim being to prevent transmission of X-linked disease. Analyses were initially performed by PCR amplification of Y chromosome-specific genomic targets. This then progressed to fluorescence in-situ hybridisation (FISH) analysis of both sex chromosomes (Delhanty, 1994). Subsequently, PCR analysis was applied to screen for single gene disorders such as cystic fibrosis (Simpson *et al.,* 2019). While the use of preimplantation genetic diagnosis (PGD) in families with known genetic disorders has been ongoing for several decades, the more recent adoption of molecular technologies capable of providing chromosome-level analysis initiated a new era of preimplantation genetic screening (PGS) (Geraedts & Sermon, 2016) Specifically, it was the use of array comparative genomic hybridisation (aCGH) that enabled genome-wide analysis of target DNA sequence that harboured copy number gains and losses. Contrastingly, linkage mapping using microsatellite markers allowing the investigation of the inheritance of linked alleles in a pedigree (Handyside *et al.,* 2010).

1.5.2 Whole genome amplification

When performing genome-wide molecular analysis, the required starting mass of DNA significantly exceeds the material available in a single cell (approximately 6 pg of DNA per nucleus); it is therefore necessary to first perform whole genome amplification (Gillooly *et al.,* 2015).

The earliest methods for whole genome amplification were developed for studies where there was a limited starting mass of available DNA (Lovmar & Syvänen, 2006). They included Degenerate Oligonucleotide Primed Polymerase Chain Reaction (DOP-PCR) which was a method created for the general amplification of targeted loci (Telenius *et al.,* 1992) and later adapted for whole genome amplification of low input DNA (Arneson *et al.,* 2008). DOP-PCR uses a primer with a fixed 5′ sequence and a degenerate six base sequence at its 3′ end. Initially, amplification is performed by low-temperature priming with extension at a raised temperature. After 5-8 cycles of amplification from the template DNA, an enrichment PCR is performed using a primer that binds the fixed 5′ sequence of the original primer. The DOP-PCR method has been optimised for commercial manufacture, with several adaptions being incorporated into the workflow as it has evolved. The Picoplex® assay (Takara Bio, USA), for example, reduces amplification bias by forming hairpin loops from the fixed 5′ sequence to prevent exponential re-amplification.

An alternative whole genome amplification strategy, Multiple Displacement Amplification (MDA), uses the strand displacement properties of the high-fidelity polymerase Phi29. Random exo-nuclease resistant primers are annealed to the template DNA and then extended using Phi29 at 30°C (Dean *et al.,* 2002). As the enzyme progressively extends from the primer, along the template molecule, it displaces any downstream primers to continue extending the amplicon. A reported benefit of this method is the preservation of sequence integrity during amplification, with a reported error rate of 10^{-5}(Esteban, Salas and Blanco, 1993).

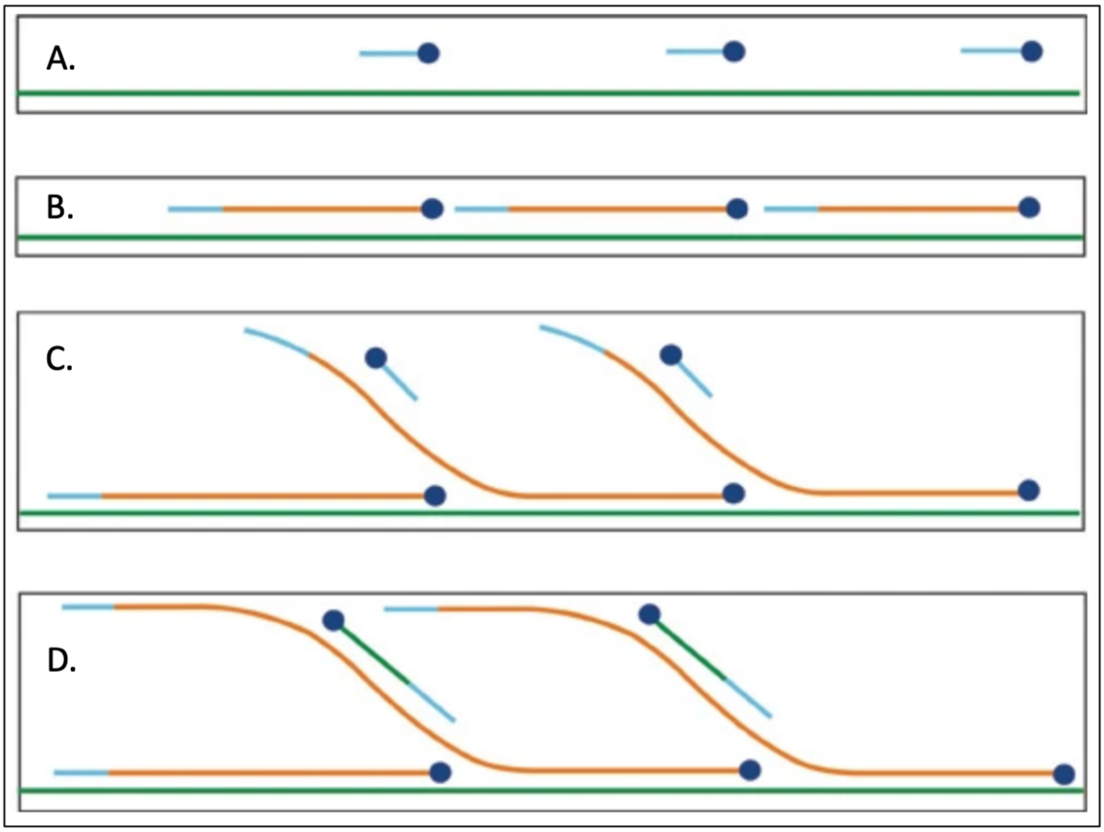

Figure 1.6: Diagram of multiple displacement amplification. (A) Random hexamers (blue line) bind to the denatured DNA (green line). (B) Phi29 DNA polymerase (blue circle) extends the primers until it reaches newly synthesized double-stranded DNA (represented by an orange line). (C) The enzyme proceeds to displace the strand and continues the polymerization, while primers bind to the newly synthesized DNA. (D) Polymerization starts on the new strands, forming a hyperbranched structure. Image adapted from Spits et al, 2006.

Most recently, Multiple Annealing and Looping-Based Amplification Cycles (MALBAC), has been developed. This approach combined from both the MDA and DOP-PCR workflows (Zong *et al.,* 2012). Primers that contain a fixed 27 nucleotide 5′ sequence and a random 8 nucleotide 3′ sequence are used in a limited strand-displacement amplification reaction comprising approximately 5 cycles of amplification. During this period semi-amplicons are first generated and are themselves amplified during the second cycle; this results in full amplicons which have complementary looped ends. The loops prevent further strand-displacement amplification of the fragments. A final PCR using a primer that is complementary to the fixed 5′ sequence amplifies the looped amplicons.

1.5.3 Context of sequence-based assays for preimplantation genetic screening

While WGA followed by aCGH has been used to successfully detect whole chromosome aneuploidies from developing embryos, at the time this research was conducted, few investigators had developed workflows for the assessment of amplified material by short-read next generation sequencing.

1.6 Considerations related to the use of short-read sequencing.

Since scientists first began to characterise the genomes of complex organisms it has been understood that differences in underlying genomic architecture affects the performance of the chosen experimental approach; the analysis of human genomes by short-read sequencing is no exception.

It is recognised that regions that are difficult or refractory to analysis can be divided into two groups; those regions for which no data is generated because the sequencing is inherently problematic and those regions where sequence data has been generated but the sequence reads are too short to be unambiguously aligned to the reference sequence (Ebbert *et al.,* 2019)

Genomic regions that are inherently difficult to sequence include those with a high GC- or AT-content. High GC-content has been associated with the first exon of some genes making comprehensive screening of disease-associated genes challenging. For individual reactions the addition of DMSO is reportedly effective for templates with up to 60% GC content (Kieleczawa, 2006). Nevertheless, for high-throughput target enrichment and sequencing workflows that have enabled the genomic footprint of a given assay to be significantly expanded, there is limited scope to optimise and manipulate the experimental parameters. While the mean GC-content of the human genome is 41% (Lander *et al.,* 2001), this contrasts significantly with the unicellular parasite *Plasmodium falciparum* that causes malaria and which has an extremely low GC (and correspondingly high AT) content of 20% (Musto *et al.,* 1997). Such organisms are particularly challenging to sequence as instruments and workflows are typically optimised for human sequencing.

Repetitive sequences (either multi-nucleotide repeats or homopolymer tracts) represent a further class of difficult to sequence regions. Errors incorporated as part of library preparation (*e.g.* polymerase slippage during PCR amplification) or noise from the sequencing process (*e.g.* when the pace of DNA synthesis within a cluster is inconsistent) can contribute to difficulties assaying these regions (Jia *et al.,* 2024).

In contrast to regions that sequence poorly and for which no (or a limited number) of sequence reads are generated, there are regions where read data is generated but it is not possible to unambiguously assign the sequence read to the reference genome. These regions are either large contiguous tandem repeats (*e.g.* centromeres, telomeres, or other short tandem repeats) or segmental duplication of specific loci that appear either in tandem with the originating locus or have been inserted into a separate region of the genome. Two types of duplicated genes exist; those that remain transcriptionally and translationally active (such as heat-shock proteins) and those that are duplicated then inactivated (becoming pseudogenes). When a bioinformatic alignment algorithm cannot unambiguously map a sequence read to the reference, it will typically position the read, at random, to one of the regions and assign it a low mapping quality score. For the widely adopted alignment program BWA the reserved mapping quality value of 0 is used; downstream data processing programs will, by default, exclude alignments with a MAPQ 0 from further analysis (Li & Durbin, 2009).

1.6.1 Medically relevant loci refractory to analysis

The systematic assessment of genomic regions that are challenging to analyse has been investigated by both diagnostic scientists seeking to understand the sensitivity of available assays and organisations that generate publicly available reference materials. This has led to comprehensive lists of genomic intervals and associated genes that are susceptible to incomplete molecular genetic analysis (Mandelker *et al.,* 2016). For clinicians and scientists that routinely analyse a set of core genes the ability to familiarise themselves with particular nuances of the genome is more straightforward than when faced with a whole exome, or whole genome, dataset.

The difficulty of challenging loci was highlighted by our own laboratory when an apparently straightforward single exon deletion of *TMEM231* exon 4, identified by comparative read-depth analysis, could not be validated. Investigating the locus by long-range PCR in combination with long-read nanopore sequencing revealed the apparent deletion was caused by a gene conversion event between *TMEM231* and its non-functional downstream pseudogene (Watson *et al.,* 2020).

In addition to tackling the analysis of pseudogene loci by long-read sequencing, some investigators have attempted to screen homologous sequences by masking pseudogene loci in the reference sequence (thereby forcing reads to the non-pseudogene locus and eliminating MAPQ 0 scores). Both approaches were used successfully to identify variants in *HYDIN*, a gene associated with primary ciliary dyskinesia and whose pseudogene, *HYDIN2,* shares 98% sequence identity. The authors identified candidate pathogenic variants in 29 individuals from 17 unrelated families. Analysis using a reference masking approach meant heterozygous variants were identified with an allele variant fraction of ~0.25 rather than (this accounts for the three wild type allele and one non-reference mutant allele). The primary analysis study was supplemented with long-read sequencing which identified a further two copy number variants (deletions of exons 36 and 17 respectively) that were refractory to detection using the reference-masking approach (Fleming *et al.,* 2024).

While pseudogenes are copied genes that have accumulated mutations to make them non-functional, they are classed as either processed (intronless) or non-processed (those that have retained their introns). Processed pseudogenes are retrotranscribed from mRNA and integrated into the host genome. Of the more than 18,000 identified pseudogenes in the human genome, approximately two thirds are classified as having been processed (Chugh, 2024).

Like non-processed pseudogenes, interference from processed pseudogene sequences can also affect the sensitivity of diagnostic testing. Our laboratory has previously characterised the integration site of the polymorphic *SMAD4* processed pseudogene. Mutations in *SMAD4* are associated with juvenile polyposis syndrome and combined juvenile polyposis/hereditary haemorrhagic telangiectasia syndrome, prompting its frequent screening and inclusion on targeted hybridisation capture panels. This enabled us to estimate the carrier frequency for the processed pseudogene-containing allele which we estimated to occur in approximately 1 in 400 people (Watson *et al.,* 2017). Visualisation of processed pseudogene reads, in a genome viewer, is defined by their gapped alignments across exon-exon junctions.

1.7.1.1 PMS2 as an exemplar challenging locus

With many pseudogene loci having been described, one such locus which is of medical importance is located at 7p22, intersecting the 15 exon *PMS2* gene (Hayward *et al.,* 2007). Molecular analysis of *PMS2* is technically challenging, due to the presence of numerous pseudogenes; 14 of these are located on the chromosome 7q arm and contain partial or complete copies of *PMS2* exons 1-5. Most problematic for assays targeting *PMS2* is the pseudogene *PMS2CL*. This is a 16 kb duplication of the 3′ end of *PMS2*, which is inverted and located approximately 700 kb away from *PMS2* towards the centromere of chromosome 7 (Bouras *et al.,* 2024). *PMS2CL* has high sequence identity to *PMS2* exons 9 and 11-15 (i.e. it does not include *PMS2* exon 10) (De Vos *et al.,* 2004).

The homology between *PMS2* and *PMS2CL* impacts the detection and classification of variants detected by hybridisation capture and short read sequencing due to the difficulty of unambiguously aligning sequence reads to the reference genome. Alternative methods have been proposed to specifically target *PMS2*, but these are not without their own limitations. Long-range PCR specific to *PMS2* can be performed using an exon 10 primer binding site which is absent from *PMS2CL* (Hayward *et al.,* 2007); this approach has been described in combination with short read sequencing (Gould *et al.,* 2018). Nevertheless, it adds costs and complexity to the testing pathway. Furthermore, restricted primer binding site locations and the limited maximum length of amplification products that can be generated from long-range PCR DNA polymerases, means confirmation and characterisation of structural or complex variants may not be possible. Further complicating the accurate interpretation of *PMS2* variants, is the presence of “hybrid alleles”; apparently pseudogene-specific variants, which due to gene conversion or reciprocal crossover events, are also be found within the *PMS2* gene itself (Ganster *et al.,* 2010). Not only can hybrid alleles complicate interpretation, but they can also affect the screening assay used to generate the data (*e.g.* due to gene conversion events taking place at primer binding sites). To aid interpretation of DNA-based results it is often helpful to assess PMS2 by immunohistochemistry.

1.7.1.2 The role of PMS2 in Hereditary Non-Polyposis Colorectal Cancer

PMS2 is one of four mismatch repair genes (MMR) for which heterozygous pathogenic mutations result in the autosomal dominant hereditary predisposition to cancer. The additional three genes include *MLH1*, *MSH2* and *MSH6*. Originally termed Hereditary Non-Polyposis Colorectal Cancer (HNPCC), due to it being the most common form of hereditary colorectal cancer (Rebuzzi, Ulivi and Tedaldi, 2023), the disease was renamed Lynch syndrome to better highlight the additional cancers with which it is associated. In addition to colorectal cancer (the second most common cause of cancer-related mortality worldwide (Sung *et al.,* 2021), Lynch syndrome is strongly associated with endometrial, ovarian, and urothelial cancer (Lindner *et al.,* 2021), as well as cancer of the stomach, small bowel, biliary tract, brain, skin, pancreas, and prostate (Idos & Valle, 2004). The prevalence of Lynch syndrome in the UK general population is estimated to be 1 in 450 (Edwards & Monahan, 2022) and Lynch syndrome is estimated to account for 3-5% of all diagnoses of colorectal cancer (Rebuzzi *et al.,* 2023). Further mechanisms of Lynch syndrome pathogenesis include a 3’ deletion of the *EPCAM* gene which results in epigenetic silencing of *MSH2* or methylation of the *MLH1* promoter.

The MMR pathway functions to repair erroneous insertions, deletions and substitutions of bases that are incorporated into the genome during DNA replication and recombination. Germline inactivation of one MMR gene allele increases the probability of complete inactivation of the protein following somatic inactivation of the remaining wild-type copy. Abnormal mismatch repair results in an accumulation of DNA replication errors, particularly in short repetitive microsatellite sequences. The identification of microsatellite instability and, or loss of MMR proteins detectable by immunohistochemistry are two hallmarks of Lynch syndrome that can aid the identification of patients prior to genetic testing.

Lynch syndrome at the cellular level is impacted by two protein families, the MutS and MutL homologues. The MutS proteins are heterodimers comprising MSH2 and MSH6 (MutSα) or MSH2 and MSH3 (MutSβ). The roles of these complexes are distinct, with MutSα detecting single base pair mismatches and smaller insertion-deletion variants, whereas the MutSβ complex recognises loop-out errors of more than two base pairs. The binding of the MutS heterodimer signals the site of mispairing, recruiting the MutL complex. The MutL dimer comprises MLH1 and one of PMS2, PMS1 or MLH3. Bases between the mismatch and an adjacent nick in the DNA are removed by exonuclease 1 (EXO1) before the strand is re-synthesised and therefore repaired by DNA polymerase β (POLB) (Olave & Graham, 2022; Yao & O'Donnell, 2012)

1.7.1.3 The Lynch syndrome clinical pathway

The lifetime risk of cancer in for Lynch syndrome patients is affected by the gene in which the pathogenic variant has arisen; carriers of *MLH1* and *MSH2* mutations have a higher cancer risk and younger age at diagnosis compared to patients with mutations in *MSH6* and *PSM2* (Stjepanovic *et al.,* 2019). Furthermore, compared to sporadic colorectal cancer, Lynch syndrome patients have an accelerated adenoma-carcinoma sequence (thought to be due to the MMR deficiency which results in the secondary accumulation of mutations in tumour suppressor and oncogenes). Consequently, sporadic colorectal cancer typically takes 10 years to develop whereas Lynch syndrome related colorectal cancer takes two years. Patients who are homozygous or compound heterozygous for pathogenic MMR gene variants present with café-au-lait spots and childhood-onset tumours; this phenotype is referred to as constitutional or biallelic MMR deficiency (CMMRD) (Wimmer *et al.,* 2014). Identifying pathogenic germline variants is therefore of considerable clinical importance.

The prevention and early detection of Lynch syndrome related cancers can increase survival in genetically confirmed patients with the precise surveillance protocol being personalised according to the affected gene and family history. Periodic colonoscopy allows the resection of polyps for the identification of early-stage colorectal cancer. For *MLH1*, *MSH2* or *EPCAM* mutation carriers this is recommended every 1-2 years beginning at the age of 20-25 years, for *MSH6* and *PMS2* mutation carriers' surveillance is recommended from aged 30-35 years (Wimmer *et al.,* 2014). For gynaecological surveillance annual transvaginal ultrasound, serum CA-125 testing and endometrial biopsy can be considered from 30-35 years. Risk reducing surgery involving bilateral salpingo-oophorectomy is an option that can be considered and should be based on whether childbearing is complete, menopausal status, comorbidity, family history and the implicated Lynch syndrome gene (there is currently insufficient evidence for this surgery to be recommended to *MSH6* and *PMS2* carriers) (Rebuzzi, Ulivi and Tedaldi, 2023). Finally, the uncertain risk of breast cancer in Lynch syndrome patients means that enhanced screening for breast cancer is not recommended but should be evaluated based on family history.

The use of aspirin and other non-steroidal anti-inflammatory drugs has been widely studied; among regular users, epidemiologists have identified a significant reduction in cancer. For people with Lynch syndrome the CAPP2 randomised trial reported a 63% reduction in colorectal cancer (and other cancers associated with condition) among those who took a daily 600 mg dose of aspirin for at least 2 years (Burn *et al.,* 2011). The ongoing CaPP3 trial (for which recruitment was closed in 2019) will investigate the effect of aspirin dose on cancer prevention. Three thousand Lynch syndrome carriers have been randomised to receive 600 mg, 300 mg or 100 mg of enteric coated aspirin daily for 2 years; they will then be followed for a minimum of 5 years taking open label 100 mg aspirin daily. The trial aims to address concerns of over possible adverse events; these include gastrointestinal bleeds or ulcers requiring hospital treatment and in very rare circumstances intracranial bleeds. The study is yet to report its findings.

Finally, recent advances in the treatment of cancer have revealed the effective use of PD-1 inhibitors in the treatment of Lynch syndrome patients with MMR deficient colorectal cancer. In the absence of a PD-1 inhibitor, the interaction between PD-L1 expressed by tumour cells and PD-1 expressed by T-cells allows the immune system to be evaded (Yu *et al.,* 2023).

1.7.2 Assay design and variant interpretation as a source of reduced test sensitivity

While test sensitivity can be reduced due to experimental constraints from the underlying genomic architecture, or bioinformatics challenges, it can also be limited by assay design. For non-whole genome sequencing assays, target genomic intervals typically comprise the coding sequences and their immediate flanking regions of selected genes. While this is a cost-effective strategy that focuses analysis on genomic regions that harbour the majority of pathogenic variants, it represents an incomplete screen of the disease-associated gene. Furthermore, clinical laboratories typically rely on the American College of Medical Genetics framework for the interpretation of identified sequence variants; these offer limited support for cryptic splicing or regulatory variants (Richards *et al.,* 2015).

With continuing updates to existing *in silico* prediction algorithms, and the creation of new software tools, the ability to comprehensively analyse whole-gene regions has significantly improved in recent years. The publication of recommendations for the clinical interpretation of variants identified in non-coding regions of the genome has supported these analyses (Ellingford *et al.,* 2022). Exemplar tools include UTRannotator, which (among other classes of UTR variant) can identify novel upstream start sites (uAUG) (Zhang *et al.,* 2021); this was the identified mechanism of action for the likely pathogenic c.-272G>A variant identified in a patient with Neurofibromatosis type 1 (Evans *et al.,* 2016). While identified pathogenic variants are frequently unique to an individual family, and require extensive functional validation, it is notable that the *COL6A1* c.930+189C>T (NM_001848.3) (chr21(GRCh37):g.47409881C>T) variant, which results in a dominantly acting splice-gain event disrupting the critical glycine repeat motif of a triple helical domain, was identified in 27 unsolved patients (Cummings *et* *al.,* 2017). (This was approximately 25% of patients clinically thought to have a collagen VI dystrophy in whom prior genetic testing was negative.) The variant's location in a CpG dinucleotide context was thought to account for its recurrent *de novo* occurrence.

Increased sequencing of both "disease free" population cohorts and patients referred to medical testing laboratories has transformed the availability of allele frequency data. Large-scale datasets are now integrated into user-friendly genome browsers such as gnomAD that can be queried from any location by genome analysts (S. Chen *et al.,* 2024). Similarly, the sharing of clinical grade variant information has been facilitated by initiatives such as ClinVar (Landrum *et al.,* 2025). Nevertheless, these datasets remain biased to communities and populations living in wealthier nations; this has several effects on the ease of variant interpretation. Firstly, the ease of rapidly identifying a variant that has been previously reported as pathogenic or likely pathogenic by another medical laboratory. Increasingly often, variant classification data is supported by additional clinical descriptions, functional test results or segregation data that can be of significant value when a variant is identified by a subsequent laboratory. Secondly, the deprioritisation of variants typically relies on a variant allele frequency being observed too frequently for the prevalence of the disease; in this situation the variant is not investigated further, allowing scientists to focus their efforts on more meaningful and productive activities. For nations or immigrant communities in which limited DNA sequencing has been undertaken, the capability to classify variants as benign or likely benign is reduced. While projects such as the Human Pangenome Reference Consortium are working to reduce this inequity by generating a reference genome sequence more representative of human genetic diversity (Liao *et al.,* 2023), it may be a number of years until this approach is adopted into mainstream diagnostic testing. Indeed, over the course of this project the transition locally from reference genome build GRCh37 to GRCh38 for diagnostic purposes happened some years after the GRCh38 release, highlighting the difficulty in enacting these changes in a clinical setting where rigorous testing and validation work is required.

1.8 Aims

The research reported in this thesis is unified by the investigation and assessment of emerging DNA sequencing workflows and technologies. Novel technologies can impact diagnostic yield, ultimately supporting their adoption into clinical practice.

Three distinct subject areas have been pursued; overarching aims included:

(i) Investigate how whole genome sequencing of single cells can provide an alternative to preimplantation genetic screening of embryos.

(ii) Determining whether linked-read technology can be deployed as an alternative library preparation methodology for the characterisation of pseudogene loci.

(iii) An assessment of the clinical utility of long-read nanopore whole genome sequencing in patients with an incomplete molecular diagnosis.

2. Materials and Methods

2.1 Specimens

2.1.1 Patient samples

Patients were referred for testing through the Yorkshire Regional Genetics Service/North-East and Yorkshire Genomics Laboratory Hub (https://ney-genomics.org.uk). Blood samples were collected from patients and family members following informed consent. Ethical approval was provided by the Leeds East Research Ethics Committee (project number 17/YH/003).

2.1.2 Cell lines

Lymphoblastoid cell lines were obtained from the Coriell Institute for Medical Research (https://www.coriell.org). Specimen GM12878 is the reference genome analysed by the Genome In A Bottle Consortium (https://www.nist.gov/programs-projects/genome-bottle) that has been extensively characterized using numerous orthogonal sequencing technologies (Zook *et al.,* 2019). A publicly available list of validated genotypes can be downloaded from the project website. Specimen AG16360 is an immortalized cell line, also obtained from the Coriell Institute. The sample was from a newborn donor that had multiple congenital abnormalities, E. coli sepsis, atypical facial features, and a possible immunodeficiency. Consistent with their clinical presentation, a diagnosis of Down Syndrome (OMIM: 190685) was confirmed following the identification of a 47, XY, +21 karyotype.

Cells were cultured in Roswell Park Memorial Institute Medium 1640 with GlutaMAX supplement and 15% foetal calf serum at 37 ^{o}C under 5% CO_2. Cells were grown to a density of $8x10^4$ cells/ml in suspension culture in T75 tissue culture flasks. In preparation for DNA extraction, the cells were transferred to a 50 ml universal tube and centrifuged to form a pellet before the supernatant was removed. The cell pellet was resuspended in phosphate buffered saline and transferred to a 2 ml sample tube compatible with the extraction instrument.

2.2 Nucleic acid extraction

2.2.1 DNA extraction from peripheral blood

Linked-read sequencing samples 1, 3 and 6, and the long-read WGS sample, which had all previously been referred for standard-of-care testing, were extracted from peripheral blood using a Chemagic 360 (Perkin Elmer) as per the manufacturer's protocol. Briefly, cells were lysed before magnetic beads were bound to the nucleic acids and removed from the solution using an automated magnetic rod. The nucleic acid bound beads were subsequently washed before the DNA was eluted (https://www.revvity.com/gb-en/product/chemagic-360-2024-0020). Sample 7 used in the linked-read experiment was extracted using the manual salting-out method described in Longmire *et al.,* 1987. For all other samples referred by external laboratories, the extraction method was unknown.

2.2.2 DNA extraction from tissue culture cell lines

DNA was isolated from a resuspended cell pellet using an EZ1 automated extractor in combination with the EZ1 DNA Tissue Kit (Qiagen, Germany). Initially, 190 µl of buffer G2 and 10 µl of Proteinase K (Qiagen) were added to the resuspended cells. The solution was incubated at 55 ^{o}C for 3 hours to allow complete cell lysis; the tube was then transferred to the EZ1 robot. Automated DNA extraction was carried out according to the manufacturer's protocol, using the EZ1 DNA Buccal Swab program card (Qiagen, 2022). The isolated DNA was eluted in 100 µl of resuspension buffer.

2.3 Nucleic acid quantification and the assessment of fragment size distributions

2.3.1 Qubit^{TM} DNA Broad Range (BR), DNA High Sensitivity (HS) and RNA Broad Range (BR) assay

To perform low-throughput quantification, the Qubit^{TM} system was used. Two assays were available; their choice of use was dependent on the sample's initial concentration. The DNA broad range assay provided a detection range of 4-2,000 ng, the DNA high sensitivity assay provided a detection range of 0.1-120 ng and the RNA broad range assay provided a detection range of 20-1000 ng per measurement. A master mix solution was first prepared, comprising 1 µl Qubit dye to 199 µl Qubit buffer. In proprietary Qubit tubes, 190 µl of the prepared master mix was combined with 10 µl pre-diluted standard, with two different standards being measured, allowing the formation of a standard curve. Typically, 1 µl of sample was quantified by combining it with 199 µl of the prepared master mix. The tubes were vortexed briefly before being incubated at room temperature for two minutes and then assayed using a Qubit^{TM} fluorometer.

2.3.2 Quant-iT Assay^{TM}

To quantify multiple DNA samples concurrently, the Quant-iT^{TM} system was used. This is specific for double stranded DNA in the 4-1000 ng range (Thermo Fisher Scientific). A standard curve was constructed by adding 2 µl each of 8 pre-diluted λ DNA standards to individual wells in column 12 of a Corning 96-well black plate. 2 µl of each sample was mixed with 198 µl of a master mix solution (1 µl Quant-iT^{TM} dsDNA BR reagent and 199 µl Quant-iT^{TM} dsDNA BR buffer) in an empty well. Fluorescence was measured using a FLUOstar Omega microplate reader, which also calculated the sample's DNA concentration using the fluorescence of the control samples in column 12 to create a calibration curve.

2.3.3 TapeStation

Automated high-throughput gel electrophoresis was performed with a 4200 TapeStation System (Agilent Technologies) to assess each sample's DNA fragment profile in the range: 35-5000 bp, as well as its molar concentrations. The selection of a standard (0.1-50 ng/µl) or high sensitivity (10-1000 pg/µl) assay was dependent on the volume and concentration of each sample. Reagents were first equilibrated to room temperature for 30 minutes and vortexed to mix, before being added to the samples in a 0.2 ml strip tube: Table 2.1 lists the volumes used for each assay type.

Table 2.1: Volume of analyte or ladder combined with sample buffer for respective TapeStation assays.

Assay	Analyte/Ladder (µl)	Sample buffer (µl)
D1000 DNA ScreenTape	1.0	3.0
D1000 DNA HS ScreenTape	2.0	2.0
D5000 DNA ScreenTape	1.0	10.0
RNA ScreenTape	1.0	5.0

The prepared strip tube was vortexed at 2,000 rpm for 1 minute, then briefly spun and loaded onto the instrument. The run was initiated using the 4200 TapeStation Controller Software (v5.1) before size distributions and concentrations were calculated using the 4200 TapeStation Analysis Software (v.4.1.1).

2.3.4 Femto Pulse

Automated pulsed-field capillary electrophoresis was performed using the Femto Pulse System (Agilent Technologies). A sample's size distribution between 300 bp and 165 kb was determined using the genomic DNA kit (FP-1002-22). Briefly, fresh gel-dye mix and conditioning solutions were prepared, and the instrument capillaries were conditioned using the 20-minute conditioning protocol. Samples were diluted to approximately 0.5 ng/µl, of which 2 µl was added to a plate that contained 18 µl of the size marker in a separate well. A ladder for fragment sizing was diluted and run in a separate well. The size separation was run using the "FP-1002-22 - gDNA 165Kb.mthds" protocol in the Femto Pulse Controller Software (v.2.0.0.3). Sample fragment distribution profiles were then calculated using ProSize data analysis software (v.4.0.2.7).

2.4 Bead-based DNA purification

Bead-based DNA purification was performed in a number of workflows using a protocol based on the solid-phase reversible immobilisation (SPRI) of DNA to magnetite polystyrene beads coated with carboxyl molecules. These beads reversibly bind DNA in the presence of polyethylene glycol (PEG) and salt. The relative concentration of PEG determines the size range of DNA molecules bound to the beads and so controls the size of DNA fragments bound and/or eluted from the bead suspension (DeAngelis, Wang and Hawkins, 1995). When beads were not included in a kit, AMPure XP (Beckman Coulter, USA) beads were used as outlined below.

Beads were first equilibrated to room temperature for 30 minutes. Usually, a 1:1 mixture of SPRI beads to target DNA solution was mixed either by vortexing (short-read applications) or inverted and placed on a Hula mixer (long-read applications), followed by incubation at room temperature for up to 15 minutes. Beads were aggregated by placing them on a magnetic stand for 5 minutes, after which the supernatant was carefully removed and discarded. While still on the magnetic stand, the beads were washed twice with 200 µl of freshly prepared 70% ethanol for 30 seconds. After removal of the ethanol solution, the beads were removed from the magnetic stand and typically air-dried for up to 5 minutes at room temperature before the addition of the required elution buffer. The beads were then incubated at room temperature for 2 minutes before the beads were aggregated by placing the suspension on the magnetic stand for 5 minutes. The supernatant containing the DNA was then removed and retained for later use.

2.5 Whole genome amplification of genomic DNA

To increase the number of genome copies within a sample, a Phi29 whole genome amplification protocol was developed by first diluting a concentrated genomic DNA and then amplifying the diluted sample. This allowed the amplified sample to be compared to the original sample to determine the utility of the amplified sample. The developed amplification protocol is as follows:

Genomic DNA samples were first diluted to a concentration of 1 ng/µl. In a 0.2 ml PCR tube, 4 µl of nuclease-free water was mixed with 1 µl of the diluted DNA and 1 µl of 500 µM Thermo Scientific Exo-Resistant Random Primer (Thermo Fisher Scientific, USA). The random primers had a 3'-terminal PTO modification that prevented the Phi29 enzyme from breaking them down with exonuclease. Using a thermal cycler, samples were denatured at 95^{o}C for 2 minutes before rapidly cooling to 30^{o}C. 5 µl of a master mix that consisted of 1 µl of Phi29 polymerase, 1 x Phi29 reaction buffer, 1 µl of BSA (0.2 mg/ml) and 1 µl of dNTP mix (20 mM) (New England Biolabs, USA) was added to each sample. The samples were incubated at 30^{o}C for 6 hours, followed by a 10-minute incubation at 65^{o}C to denature the Phi29 polymerase. All reactions were performed in triplicate. Negative controls were created by substituting either the DNA sample or polymerase with the equivalent volume of nuclease-free water.

2.6 Microsatellite genotyping

Microsatellite PCR was used to verify the amplification of DNA samples by MDA. Publicly available genotype data for the GM12878 sample was used to find heterozygous microsatellites that could be amplified with primers in the ResGen Map Pairs Human Screening Set (version 10) from Invitrogen, USA. Six markers, as outlined in Table 2.2, were selected for PCR optimisation and genotyping using a fluorescently labelled primer pair. Each amplification reaction comprised 1 µl of MDA-amplified genomic DNA (5 ng/µl), 0.5 µl of forward primer (10 µM), 0.5 of reverse primer (10 µM) and 10 µl of Megamix PCR master mix (Microzone Ltd, Stourbridge, UK). Thermocycling conditions comprised 95^{o}C for 3 minutes, followed by 35 cycles of 95^{o}C for 30 seconds, 55^{o}C for 30 seconds and 72^{o}C for 1 min, before a final extension at 72^{o}C for 5minutes. To resolve the fluorescently labelled amplicons, each PCR reaction was diluted 1 in 20 using nuclease-free water, with 0.5 µl transferred to a 96-well plate containing 0.25 µl of the GeneScan LIZ 500 size standard (Applied Biosystems) and 10 µl Hi-Di Formamide (Applied Biosystems). The samples were denatured at 95 ^{o}C for 30 seconds, before being chilled on ice before loading on a 3730 DNA Analyzer (Applied Biosystems). The resultant data files were then analysed using PeakHeights (http://www.insilicase.com/Desktop/PeakHeights.aspx) (Ingham *et al.,* 2013).

Table 2.2: Microsatellite primers for heterozygous genotypes identified in cell line GM12878.

Marker name	**Genomic position***			**Forward primer**	**Reverse primer**	**Expected size range (bp)**	**Fluorophore**
	Chr	**Start**	**Stop**				
D1S3669	1	17457172	17457352	TTTTGTTTCTTGATCTGGGC	TGTTAAACTTTTCACTGAGGTATAA	171-215	FAM
D2S1363	2	226164890	226165078	TTCTGCTTTCTCTGACTGTATCA	ATTCTTTGTCTCCCCAGTTG	172-192	HEX
D4S1627	4	44176796	44176974	AGCATTAGCATTTGTCCTGG	GACTAACCTGACTCCCCCTC	177-201	HEX
D6S1027	6	168809255	168809385	CGTTCTGCACATGTATCCTG	TGCTCTGTCTATGGAGTAGCC	110-150	TET
D7S1824	7	140312765	140312968	GCACCTGTTTGATTCAGTCA	CCAGCCTGTGTGACTATGTG	163-203	FAM
D9S934	9	118333476	118333699	TTTCCTAGTAGCTCAAGTAAAGAGG	AGACTTGGACTGAATTACACTGC	206-230	TET

Chr: Chromosome. *Genomic coordinates provided according to human reference genome build hg38.

2.7 Single nucleotide polymorphism genotyping by TaqMan^{TM} assay

To determine the allelic balance of whole genome amplified DNA samples TaqMan genotyping was performed. Amplification reactions were set up in a 384-well MicroAmp Optical Reaction plates (Applied Biosystems, USA) with each reaction comprising 2.5 µl of TaqMan Master Mix (Applied Biosystems), 0.25 µl of 20x TaqMan probe mix and 2.25 µl of DNA (5 ng/µl). Reactions were performed in triplicate using a QuantStudio 5 real-time PCR system (Applied Biosystems, USA). Catalogue TaqMan SNP Genotyping probes were used and are outlined in Table 2.3 below. Thermocycling conditions consisted of a 10-minute incubation at 95 ^{o}C then 40 cycles of 95 ^{o}C for 15 seconds and 60 ^{o}C for 1 minute. Data was processed and interpreted using the Connect Genotyping application (version 4.1) (Thermo Fisher Scientific, USA).

Table 2.3: Catalogue TaqMan SNP Genotyping Probes used for genotyping assay.

TaqMan SNP Probe Assay ID	SNP ID
C_9524069_10	rs1156253
C_1121246_10	rs747039
C_11907549_1_	rs1872575
C_11245682_10	rs6811238
C_2073009_10	rs1109037

2.8 Short-read whole genome sequencing

2.8.1 Library preparation and sequencing

To prepare the samples for sequencing, 1 µg of input material was sheared using a Covaris E220 focused ultrasonicator (Covaris Inc, USA). A fragment size distribution of 250-300bp was achieved and confirmed using the Agilent D1000 ScreenTape. NGS libraries were prepared using the NEBNext Ultra DNA Library Preparation Kit for Illumina (New England BioLabs, USA). Sheared fragments underwent end-repair, dA-addition and sequencing adaptor ligation, followed by an AMPure bead (Beckman Coulter) clean-up and an enrichment PCR, at which point unique index sequences were added to each library. After a final AMPure clean-up the samples were run on an

Agilent High Sensitivity D1000 ScreenTape to confirm the success of the library preparation. The libraries were quantified using the Quant-iT dsDNA Broad Range Assay kit (Invitrogen) and analysed on a FLUOstar Omega plate reader (BMG Labtech). Both the sizing and quantification assays identified that sample MDA4 had produced a sub-optimal library. This was therefore eliminated for sequencing purposes. The rest of the libraries were pooled at a quantity of 200 ng each for sequencing on a NextSeq 500 (Illumina Inc., USA).

The pooled libraries were diluted and denatured according to Illumina's recommendations, with a final sample loading concentration of 1.5 pM with 1% spike-in of Illumina PhiX control v3. Sequencing was performed using paired-end 151 bp reads on a high-output v2 NextSeq cartridge.

2.8.2 Data processing workflow

Data was demultiplexed using bcl2fastq Conversion Software (Illumina), which created a sample-specific gzip-compressed FASTQ formatted file. Single-end reads for the CNVseq pipeline were 76 bp in length, while paired-end reads for variant calling pipelines were 151 bp long.

The demultiplexed fastq files were aligned to the human reference genome GRCh37, which was the genome build associate with the diagnostic pathways at that time. Alignment was done using the mem function of the BWA short read aligner (Li & Durbin, 2009). The aligned reads were exported as plain text SAM formatted files were subsequently ordered by chromosomal coordinate and reformatted as BAM files using samtools' view and sort functions. Finally, the BAM files were indexed using the index function of samtools.

As part of the established CNVseq pipeline (Hayes *et al.,* 2013), the read coordinate data was extracted from the BAM alignment files extracted from the 76bp single read BAM files using samtools awk and sort functions. The number of reads in the aligned data from control and patient data were counted and used to determine a window size.

Initially, the window was set such that it, on averaged, contained 80 reads across the genomes of the patient and control files. Due to the uneven DNA distribution in the amplified samples this window was incrementally increased to 160, 500, 1000 and 2000 reads per window to reduce the noise in the reported values.

2.9 Linked-read workflow.

2.9.1 Linked-read library preparation, target enrichment and sequencing.

10 µl of denaturing agent aliquoted into an 8-tube strip tube to which 10 µl diluted DNA (1 ng/µl) was add, mixed 10 times using wide-bore tips and incubated at room temperature for 5 minutes. Following the incubation, 3 µl denatured DNA was transferred into the to 97 µl of sample master mix (Table 2.4) on ice and mixed using wide-bore tips.

Table 2.4: Components of 10X Gel Beads-in-emulsion (GEM) Sample Master Mix.

Reagent	Volume per reaction (µl)
Genome Reagent Mix	89.5
Additive A	3.0
Genome Enzyme Mix	5.0

90 µl of the sample mixture was transferred to row 1 of the 10X Genome Chip. A Gel Bead Strip was placed into a 10X Vortex Adapter and vortexed for 30 seconds. The foil seal was punctured and 85 µl of Genome Gel Beads was aspirated and slowly dispensed into row 2 of the Genome Chip. 270 µl of Partitioning Oil was added to the wells in row 3 of the Chip and the Chip was covered with a 10X Gasket and transferred to the Chromium^{TM} Controller for Gel Bead-in-emulsion (GEM) generation using the "Genome" program. Once generated, 125 µl GEMs were slowly retrieved from the Recovery Wells of the plate and transferred to a PCR plate. The plate was incubated on a thermal cycler at 30^{o}C for 3 hours followed by 10 minutes at 65^{o}C.

Following GEM incubation, 125 µl of Recovery Agent was added to each well and mixed. The entire volume was transferred to a strip tube and vortexed for 15 seconds to separate the liquid into phases. The lower pink phase was removed and discarded, retaining the upper aqueous phase containing the sample.

Table 2.5: Components of 10X DynaBeads GEM Clean-up Mix

Reagent	Volume per reaction (µl)
Buffer Sample Clean Up 1	130.0
DynaBeads MyOne Silane	14.0
Additive A	6.0

150 µl of the DynaBeads Cleanup Mix (Table 2.5) was added to each sample, pipette mixed thoroughly and left at room temperature for 10 minutes. After 10 minutes the strip tube was transferred to a 10X^{TM} Magnetic Separator in the high position. Once the solution was clear the supernatant was removed and discarded. Next, 250 µl of fresh 80% ethanol was added and left for 30 seconds, before being removed and discarded. A further ethanol wash was performed with 200 µl of 80% ethanol. Once ethanol was removed the strip tube was briefly spun down and returned to the Magnetic Separator in the low position, and any residual ethanol removed. The strip was taken off the magnet and 51 µl of Elution Solution I (Table 2.6) was added.

Table 2.6: Components of 10X GEM Clean-up Elution Solution I

Reagent	Volume per reaction (µl)
Buffer EB	89.0
10% Tween 20	1.0
Additive A	10.0

The solution was incubated for 30 seconds before pipette mixing to resuspend the beads. The mixed beads were incubated at room temperature for 5 minutes before briefly spinning and returning to the Magnetic Separator in the low position. 50 µl supernatant sample was transferred to a new strip tube.

The SPRIselect Reagent (Beckman Coulter) was vortexed until the beads were resuspended and 60 µl was added to each sample and pipette mixed. After 5 minutes incubation at room temperature, the strip was placed on the Magnetic Separator in the High position. Once the solution was clear the supernatant was removed and discarded. 125 µl fresh 80% ethanol was added and left for 30 seconds, before being removed and discarded. A further ethanol wash was performed with 200 µl 80% ethanol. Once ethanol was removed the strip tube was briefly spun down and returned to the Magnetic Separator in the low position, and any residual ethanol removed. The strip was taken off the magnet and 52 µl of Elution Solution II, prepared as described in Table 2.7 was added with pipette mixing to resuspend the beads. The beads were incubated at room temperature for 5 minutes before briefly spinning and returning to the Magnetic Separator in the low position. 52 µl supernatant sample was transferred to a new strip tube. The process was repeated from the addition of SPRIselect beads for a total of two clean-ups.

Table 2.7: Components of 10X GEM Clean-up Elution Solution II

Reagent	Volume per reaction (µl)
Buffer EB	196.0
Additive A	4.0

The size profile of the post-GEM samples was determined using a TapeStation using an Agilent D5000 ScreenTape. 50 µl of each sample was sheared using a Covaris E220 focused-ultrasonicator (Covaris Inc) to a target fragment size of 225 bp was achieved, as confirmed using a TapeStation with an Agilent D1000 ScreenTape. An End Repair and A-tailing Mix was prepared according to Table 2.8.

Table 2.8: Components of 10X End Repair and A-tailing Mix

Reagent	Volume per reaction (µl)
Nuclease-free Water	2.5
End Repair/A-tailing Buffer	7.5
End Repair/A-tailing Enzyme	15

25 µl of this mix was added to each sheared sample. Samples were incubated on a thermal cycler for 30 minutes at 20°C followed by 30 minutes at 65°C with the lid temperature set to 85°C.

Table 2.9: Components of 10X Adaptor Ligation Mix

Reagent	Volume per reaction (µl)
Ligation Buffer	22.0
DNA Ligase	11.0
Adaptor Mix	2.5

Adaptor Ligation Mix was prepared according to Table 2.9. 35.5 µl was added to each sample and mixed by pipetting. Samples were incubated on a thermal cycler at 30°C for 15 minutes with the lid temperature set to 30°C. The SPRIselect Reagent was vortexed until resuspended and 198 µl was added to each sample and pipette mixed. After 5 minutes incubation at room temperature, the strip was placed on the Magnetic Separator in the High position. Once the solution was clear the supernatant was removed and discarded. 250 µl fresh 80% ethanol was added and left for 30 seconds, before being removed and discarded. A further ethanol wash was performed with 250 µl 80% ethanol. Once ethanol was removed the strip tube was briefly spun down and returned to the Magnetic Separator in the low position, and any residual ethanol removed. The strip was taken off the magnet and 40 µl of Buffer EB was added and pipette mixed to resuspend the beads. The beads were incubated at room temperature for 5 minutes before briefly spinning and returning to the Magnetic Separator in the low position. 40 µl supernatant sample was transferred to a new strip tube. Sample Index PCR Mix was prepared according to Table 2.10.

Table 2.10: Components of 10X Index PCR Mix

Reagent	Volume per reaction (µl)
Amplification Master Mix	50.0
Forward PCR Primer	5.0

55 µl Sample Index PCR Mix was added to each ligated sample followed by 5 µl of a unique Chromium i7 Sample Index per sample. Samples were mixed by pipetting and place on thermal cycler under PCR conditions of 98°C for 45 seconds, 12 cycles of 20 seconds at 98°C, 30 seconds at 54°C, and 20 seconds at 72°C, with a final extension of 72°C for 1 minute. Following PCR, the samples underwent a final SPRIselect bead clean-up as described previously, with 180 µl SPRIselect beads, 250 µl 80 % ethanol for the washes, and elution in 20 µl nuclease-free water. Sample size distribution was checked on a TapeStation with an Agilent D1000 ScreenTape, and libraries were quantified using the Qubit BR Assay. 187.5 ng of each library was pooled into a single well of a PCR plate and vacuum concentrated until dried in preparation for hybridisation with the Twist capture probes.

A probe solution was prepared according to Table 2.11. The mix was heated to 95°C in a thermal cycler for 2 minutes and then cooled on ice for 5 minutes before being left to equilibrate at room temperature for 5 minutes.

Table 2.11: Components of Twist PMS2/PMS2CL Hybridisation Probe Solution

Reagent	Volume per reaction (µl)
Hybridisation Mix	20.0
PMS2/PMS2CL Probes	4.0
Nuclease-free Water	4.0

The dried library pool was resuspended by the addition of 5 µl Twist Blocker Solution and 8 µl Twist Universal Blockers. The mix was heated to 95°C in a thermal cycler for 5 minutes and then cooled at room temperature for 5 minutes before the addition of the probe solution. The hybridisation reaction was mixed by pipetting and briefly spun down before incubation on a thermal cycler at 70°C for 16 hours with the heated lid set to 85°C.

100 µl of Streptavidin Binding Beads were prepared for the library capture in a 1.5 ml tube. This consisted of the addition of 200 µl Binding Buffer, pipetting to mix and placing on a magnetic stand for 1 minute. The supernatant was removed, and the beads were resuspended in 200 µl of fresh Binding Buffer. The process was repeated for a total of three bead washes.

Once the 16-hour hybridisation was completed, the entire hybridisation reaction was transferred to the resuspended Streptavidin Binding Beads and the tube placed on a rotator for 30 minutes. The tube was placed on a magnetic stand, and the supernatant removed and discarded. The beads were resuspended in 200 µl Wash Buffer 1, and the entire reaction was transferred to a fresh 1.5 ml tube. The tube was placed on a magnetic stand and the supernatant removed and discarded. The tube was removed from the magnet and the beads were resuspended in 200 µl of pre-warmed (48°C) Wash Buffer 2 and incubated at 48°C for 5 minutes. The tube was then placed on a magnetic stand, and the supernatant was removed and discarded. The beads underwent a further two washes and 48^{o}c incubations with Wash Buffer 2. After the final wash, the beads were resuspended in 45 µl of nuclease-free water.

Table 2.12: Components of Twist Post-capture PCR enrichment Mix

Reagent	Volume per reaction (µl)
Amplification Primers, ILM	20.0
KAPA HiFi HotStart ReadyMix	4.0

22.5 µl of the Streptavidin Binding Bead Slurry was added to a PCR mix (Table 2.12) and mixed by pipetting, before PCR amplification by incubation at 98^{o}C for 45 seconds, followed by 8 cycles of 15 seconds at 98^{o}C, and finally 30 seconds at 60^{o}C, and 30 seconds at 72^{o}C, with a final extension of 72^{o}C for 1 minute. Following PCR, the samples underwent a final bead purification using 50 µl of Twist DNA Purification Beads, with 2 x 80% ethanol washes and final elution in 30 µl nuclease-free water. Sample pool size distribution and molarity was checked on a TapeStation using an Agilent High Sensitivity D1000 ScreenTape.

The library pool was diluted and denatured according to Illumina's recommendations. A final loading concentration of 1.6 pM, with a 1% spike-in of Illumina PhiX control v3, was sequenced on a NextSeq 500 to produce 151 bp, paired-end reads using a high-output v2 NextSeq 500 cartridge.

2.9.2 Linked-read data processing pipeline.

The exported BCL data files were converted to patient-specific gzip-compressed FASTQ-formatted text files using the 'mkfastq' function of the longranger (v.2.2.2) (10X Genomics, Pleasanton) software. The read data was aligned to the human reference genome (build hg19) using the 'targeted' function of the longranger software to create a BAM file with alignments ordered by their genomic coordinates. A bed file was created and used in combination with samtools (v.1.12) to identify the BX tags for sequences that mapped to the unique regions of either hg19.chr7:6035165-6038906 or hg19.chr7:6729936-6769936 that contained *PMS2* exons 6-8 and a 40 kb interval located 5 kb upstream of *PMS2CL,* respectively. Subset-bam (v.1.1.0) was next used to identify and extract sequences mapping to this region, exporting the reads to a FASTQ-formatted text file. The reads were then aligned to the masked hg19 reference sequence using the ‘align’ function of longranger (Figure 2.1).

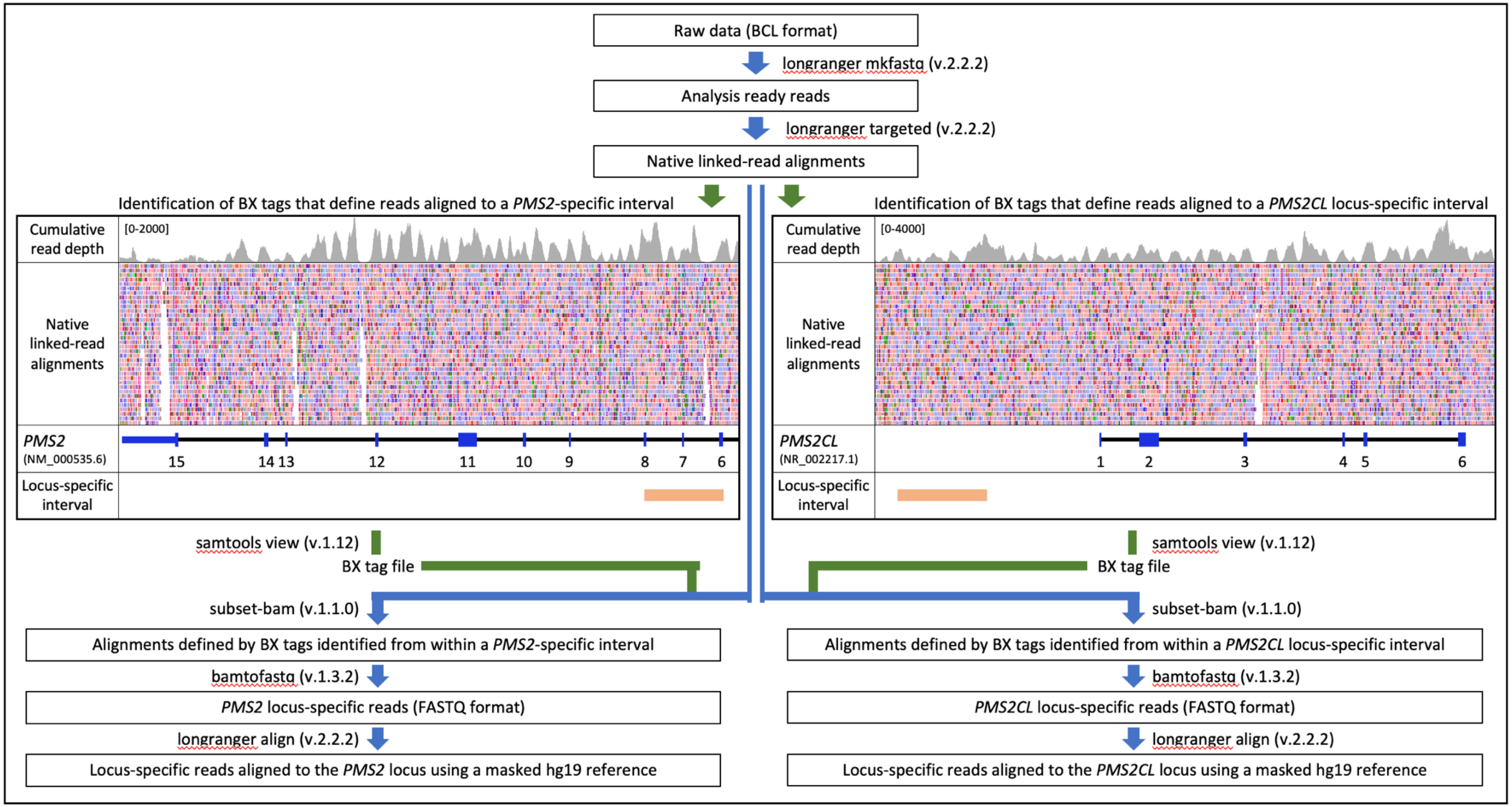
Raw data (BCL format)
longranger mkfastq (v.2.2.2)
Analysis ready reads
longranger targeted (v.2.2.2)
Native linked-read alignments
Identification of BX tags that define reads aligned to a PMS2-specific interval
Identification of BX tags that define reads aligned to a PMS2CL locus-specific interval
Cumulative read depth
[0-2000]
Native linked-read alignments
PMS2
(NM_000535.6)
15
14 13
12
11
10
9
8
7
6
Locus-specific interval
Cumulative read depth
[0-4000]
Native linked-read alignments
PMS2CL
(NR_002217.1)
1
2
3
4 5
6
Locus-specific interval
samtools view (v.1.12)
BX tag file
subset-bam (v.1.1.0)
samtools view (v.1.12)
BX tag file
subset-bam (v.1.1.0)
Alignments defined by BX tags identified from within a PMS2-specific interval
bamtofastq (v.1.3.2)
PMS2 locus-specific reads (FASTQ format)
longranger align (v.2.2.2)
Locus-specific reads aligned to the PMS2 locus using a masked hg19 reference
Alignments defined by BX tags identified from within a PMS2CL locus-specific interval
bamtofastq (v.1.3.2)
PMS2CL locus-specific reads (FASTQ format)
longranger align (v.2.2.2)
Locus-specific reads aligned to the PMS2CL locus using a masked hg19 reference

Figure 2.1: Schematic overview of the linked-read data processing pipeline for 10X data.

2.9.3 Comparison between known GIAB genotypes and 10X linked read dataset.

The high confidence bed file (HG001_GRCh37_1_22_v4.2.1_benchmark.bed) was downloaded from the GIAB website and filtered for the unique (non-pseudogene encompassing) region of the target locus. This was defined by the first interval upstream of *PMS2* (chr7:6049657-6049792 located in *AIMP2*) through to the first interval in *SPDYE20P* (chr7:6762920-6763395). *SPDYE20P* being the first gene upstream of *PMS2CL* and contained a total of 332 intervals spanning 649 kb of genomic sequence. Bcftools (v.1.14) was used in combination with the curated interval file to generate separate SNP and INDEL datasets from the GIAB benchmark VCF (HG001_GRCh37_1_22_v4.2.1_benchmark.vcf.gz) and those called by the variant function of longranger (Sample02_NA12878_10X.vcf.gz). The National Institute of Standards and Technology recommended benchmarking tool vcfeval (v.3.12.1; https://github.com/RealTimeGenomics/rtg-core) developed by Real Time Genomics was used to perform a pairwise comparison between identified variants and publicly available genotypes for the NA12878 sample.

2.10 Long-range PCR enrichment at the PMS2 locus

Long PCR products were generated for PMS2 exons 11-15 using the primers in Table 2.13. The PCR reaction was performed with SequelPrep^{TM} Long PCR Kit (Invitrogen, USA) comprising of 1 µl of sample DNA, 2 µl 10X Reaction Buffer, 0.36 µl SequelPrep Long Polymerase, 0.4 µl DMSO, 1 µl Enhancer B, 0.5 µl each 10 µM primer and 16.24 µl nuclease-free water. The reactions were placed on a thermal cycler and incubated under the following conditions: 94°C for 2 minutes, then 10 cycles of 94°C for 10 seconds, 60°C for 15 seconds and 68°C for 16 minutes, followed by a further 20 cycles of 94°C for 10 seconds, 60°C for 15 seconds and 68°C for 16 minutes with 20 seconds added to the 68°C extension per cycle, and a final 72°C extension for 10 minutes.

Table 2.13: Primers for amplification of PMS2 exons 11-15

Primer	Sequence
Forward	GATTAGAAGAAGTCTGCAGTGACTGCATT
Reverse	ACACACACGAGCGCATGCAAACATAGA

PCR products were run out on a 1 % agarose gel at 100 V for 2 hours. DNA products of the appropriate size were excised with a scalpel and extracted from the gel slice using the QIAquick Gel Extraction Kit (Qiagen, Germany) as follows: Briefly, gel slices were weighed and dissolved in 3 volume equivalents of the QG buffer by incubating at 50^{o}C until dissolved, at which point 1 volume equivalent of isopropanol was added and mixed by vortexing. The sample was then transferred to a QIAquick spin column, centrifuged at 179000 x g for 1 minute and the flow-through discarded. The column was washed with first with 500 µl QG buffer and 750 µl Buffer PE each time the wash solution was discarded following centrifugation. A final centrifuge step was conducted to remove residual buffers, after which the column was transferred to a new 1.5 ml tube and 30 µl EB Buffer was added to the column membrane. After 1 minute incubation at room temperature, the column was centrifuged to collect the eluted PCR product in the 1.5 ml tube, which was then quantified using the Qubit BR Assay.

To prepare the PCR products for sequencing, 400 ng of material was sheared using a Covaris E220 focused-ultrasonicator (Covaris Inc) to create a fragment distribution of 250-300bp which was confirmed using a TapeStation with a D1000 ScreenTape. NGS libraries were prepared using the NEBNext Ultra DNA Library Preparation Kit for Illumina (New England BioLabs). Sheared fragments underwent end-repair, A-addition and sequencing adaptor ligation, followed by an AMPure bead (Beckman Coulter) clean-up and an enrichment PCR, at which point unique index sequences were added to each library. After a final AMPure clean-up the samples were quantified on a TapeStation using a High Sensitivity D1000 ScreenTape to confirm the addition of the adaptor sequences. Libraries were quantified by Qubit BR Assay and pooled to for an equimolar sequencing pool for subsequent sequencing.

The pool was diluted to 4 nM and denatured according to Illumina’s recommendations, before being diluted to the final loading concentration of 4 pM, with a 1 % spike-in of Illumina PhiX control v3. The libraries where then sequenced on a MiSeq with a 300 cycle v2 MiSeq cartridge to produce 151 bp paired end data.

2.11 Long-read whole genome sequencing

The size profile of the DNA sample was assessed using the Femto Pulse as described in Section 2.3.4. 10 µg DNA was transferred to a Covaris g-TUBE and sheared by centrifuging at 1503 RCF for 30 seconds, inverting and centrifuging for a further 30 seconds. A short-read elimination (SRE) procedure was performed to remove as many fragments below 10kb as possible. 50 µl SRE Regular buffer (PacBio, USA) was added to the sheared DNA and pipette mixed with a wide-bore tip. The sample was centrifuged at 10,000 RCF for 1 hour. The supernatant was removed, 200 µl fresh 70 % ethanol was added and the sample was centrifuged at 10,000 RCF for 10 minutes. The ethanol wash step was repeated once. Following removal of all residual ethanol, 50 µl Elution Buffer was added to the sample pellet, which was left in the fridge overnight to resuspend. The following day, the sample was pipette mixed with a wide-bore tip to fully resuspend.

DNA was prepared for long-read nanopore sequencing using Ligation Sequencing SQK-LSK114 (Oxford Nanopore technologies, UK) as follows. 1 µg genomic DNA was combined with the end repair mix described in Table 2.14 in a PCR tube and flicked to mix.

Table 2.14: Components of End Repair Mix for Nanopore sequencing preparation

Reagent	Volume per reaction (µl)
NEBNext FFPE DNA Repair Buffer	3.5
NEBNext FFPE DNA Repair Mix	2.0
Ultra II End-prep Reaction Buffer	3.5
Ultra II End-prep Enzyme Mix	3.0

Following a brief centrifugation, the sample was incubated in thermal cycler for 10 minutes at 20°C followed by 10 minutes at 65°C. The sample was transferred to a 1.5 ml tube to which 60 µl of AMPure XP beads were added. The tube was flicked to mix and placed on a Hula mixer for 10 minutes at room temperature, after which it placed on a magnetic stand to pellet the beads, the supernatant was discarded. The beads were washed with 200 µl fresh 70 % ethanol for a total of two washes, as described previously. The DNA was eluted in 60 µl nuclease-free water and transferred to a fresh 1.5 ml tube.

Table 2.15: Components of Ligation Mix for Nanopore sequencing preparation

Reagent	Volume per reaction (µl)
Ligation Buffer (LNB)	25.0
Ligation Adapter (LA)	5.0
NEBNext Quick T4 Ligase	10.0

A ligation reaction was set up by sequentially adding the reagents in Table 2.15 and flicking the tube to mix after which the ligation reaction was incubated at room temperature for 10 minutes. The ligation mix was cleaned using 40 µl AMPure XP beads, as described earlier, except the beads were washed with 200 µl Long Fragment Buffer (LFB) and the DNA eluted in 32 µl Elution Buffer (EB).

An R10.4.1 flow cell was loaded onto the PromethION and QC'd using MinKNOW (v23.11.4) to check available pores for sequencing. A priming mix was prepared according to Table 2.16.

Table 2.16: Components of PromethION flow cell priming mix

Reagent	Volume per reaction (µl)
Flow Cell Flush (FCF)	1170.0
Flow Cell Tether (FCT)	30.0

Any residual air in the flow cell inlet port was removed prior to loading by opening the port cover, inserting a p1000 tip with the pipette set to 200 µl and slowly turning the pipette dial up towards 220 µl until a small volume of fluid was visible in the tip, at which point the tip was removed and discarded. 500 µl priming mix was added to the flow cell and incubated at room temperature for 5 minutes, followed by a further 500 µl priming mix. The sample was prepared for sequencing according to Table 2.17, with 10 fmol library in a 32 µl volume used for loading.

Table 2.17: Components of PromethION sequencing mix

Reagent	Volume per reaction (µl)
Sample Library	32.0
Sequencing Buffer (SB)	100.0
Library Solution (LIS)	68.0

The mixture was loaded onto the flow cell and a sequencing run of 72 hours was started. Due to the low yield of the initial library preparation for the TTC37 case, the majority of the sample was used for first flow cell load. To enable a nuclease flush and reload of the flow cell, a second library was prepared using 1 µg of the previously size selected DNA.

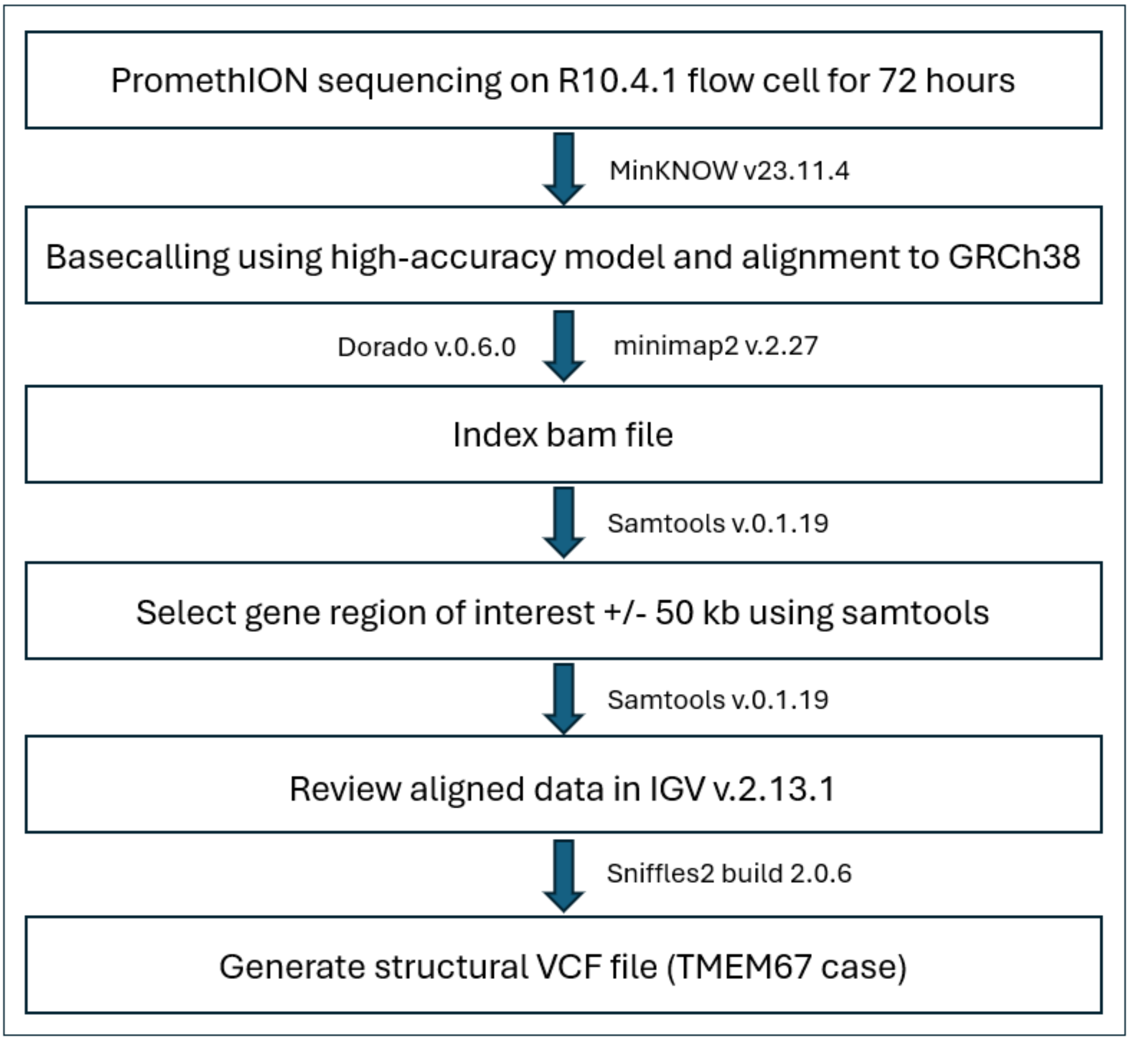
PromethION sequencing on R10.4.1 flow cell for 72 hours
MinKNOW v23.11.4
Basecalling using high-accuracy model and alignment to GRCh38
Dorado v.0.6.0
minimap2 v.2.27
Index bam file
Samtools v.0.1.19
Select gene region of interest +/- 50 kb using samtools
Samtools v.0.1.19
Review aligned data in IGV v.2.13.1
Sniffles2 build 2.0.6
Generate structural VCF file (TMEM67 case)

Figure 2.2: Overview of the data processing pipeline for long-read nanopore sequencing.

2.12 Short read whole transcriptome sequencing

Automated extraction of RNA was performed using a peripheral blood PAXgene Blood RNA kit, in tandem with a Qiagen QIAcube instrument. Blood was collected in a PAXgene tube and stored at -20°C until extraction. The blood was thawed at room temperature for two hours before centrifugation at 3000 x g for 10 minutes to pellet the nucleic acids. The supernatant was discarded and the pellet washes with 4 ml nuclease-free water. A further centrifugation step was performed, and the supernatant discarded. The pellet was resuspended in 350 µl resuspension buffer BR1 and transferred to a processing tube. This was placed in the thermo-shaker unit of the QIAcube and processed using the “PAXgene Blood RNA Part A” protocol. Once completed, the resultant RNA was returned to the thermo-shaker unit and denatured using the “PAXgene Blood RNA Part B” protocol.

Prior to short-read sequencing preparation, globin mRNA transcripts were removed using the Invitrogen GLOBINclear kit (Thermo Scientific, USA) as follows. 30 µl Streptavidin Magnetic Beads was transferred to a 1.5ml tube and placed on a magnetic rack. The supernatant was removed, and the beads resuspended in 30 µl of Streptavidin Bead Buffer. The prepared beads were stored on a hot block at 50°C until needed. 14 µl of RNA was combined with 1 µl of Capture Oligo Mix and 15 µl of pre-warmed 2X Hybridisation Buffer in a strip tube and incubated on a thermal cycler at 50°C for 15 minutes. The pre-prepared Streptavidin Magnetic Beads were then added to the sample, briefly vortexed to mix and then placed back on the thermal cycler at 50°C for 30 minutes. After the incubation, the sample was placed on a magnetic rack and the supernatant containing the depleted RNA transferred to a fresh strip tube. 100 µl of RNA Binding Buffer was added to the sample, followed by 20 µl of Bead Resuspension Mix (Table 2.18).

Table 2.18: Components of NEBNext Ultra II Directional RNA Bead Resuspension Mix

Reagent	**Volume per reaction (µl)**
RNA Binding Beads	10.0
RNA Bead Buffer	4.0
100 % Isopropanol	6.0

The strip tube was placed on the magnetic rack, the supernatant removed and 200 µl RNA Wash Solution added. The sample was vortexed briefly to mix and returned to the magnet. The wash solution was removed, and the beads allowed to air dry for 5 minutes. 30 µl of pre-warmed Elution Buffer was added to the sample, which was placed on a thermal cycler at 50^{o}C for 5 minutes. The strip tube was returned to the magnet and the sample transferred to a fresh strip tube. The globin-depleted RNA was quantified using the using the Qubit RNA BR Assay, with a concentration of 71.2 ng/µl.

1 µg of depleted RNA was taken forward for mRNA selection, cDNA synthesis and library preparation using the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina. 20 µl of oligo dT beads were added to 100 µl RNA Binding Buffer in a 1.5 ml tube and placing on a magnetic rack. The supernatant was removed and discarded, and the beads resuspended in a further 100 µl of RNA Binding Buffer. The tube was returned to the magnet and the supernatant removed and discarded before finally resuspending the beads in 50 µl RNA Binding Buffer.

The prepared oligo dT beads were added to the RNA sample which was placed on a thermal cycler at 65^{o}C for 5 minutes and then chilled to 4^{o}C, at which point it was incubated for 5 minutes at room temperature. The strip tube was then placed on a magnet stand and the supernatant discarded. The sample bound to oligo dT beads underwent two washes with 200 µl of Wash Buffer, followed by resuspension in 50 µl of Tris Buffer. The suspension was incubated on a thermal cycler at 80^{o}C for 2 minutes to temporarily release the mRNA from the beads. 50 µl of RNA Binding Buffer was then added, mixed and incubated at room temperature for 5 minutes to re-bind the mRNA. The strip tube was placed on a magnet and the supernatant removed and discarded. The RNA bound to the oligo dT beads underwent a final wash with 200 µl of Wash Buffer, with the wash supernatant removed and the beads resuspended in 11.5 µl of First Strand Priming Mix which was prepared according to Table 2.19.

Table 2.19: Components of NEBNext Ultra II Directional RNA First Strand Priming Mix

Reagent	Volume per reaction (µl)
NEBNext First Strand Synthesis Reaction Buffer	8.0
NEBNext Random Primers	2.0
Nuclease-free water	10.0

The beads were incubated for 15 minutes at 94°C and then immediately transferred to ice for 1 minute. The strip tube was placed on a magnetic rack and 10 µl of supernatant transferred to a fresh strip tube. 8 µl of NEBNext Strand Specificity Reagent and 2 µl of NEBNext First Strand Synthesis Enzyme Mix was added, and the strip tube placed on a thermal cycler and incubated as follows: 25°C for 10 minutes, 42°C for 15 minutes and 70°C for 15 minutes.

Table 2.20: Components of NEBNext Ultra II Directional RNA Second Strand Synthesis Mix

Reagent	Volume per reaction (µl)
NEBNext Second Strand Synthesis Reaction Buffer with dUTP	8.0
NEBNext Second Strand Synthesis Enzyme Mix	4.0
Nuclease-free water	48.0

A Second Strand Synthesis Mix was prepared according to Table 2.20, with 60 µl added to the sample and incubated for 1 hour at 16 °C. The synthesised cDNA underwent an AMPure bead clean-up using 144 µl of beads, two 80 % ethanol washes and elution in 50 µl of 0.1 x TE Buffer. The purified cDNA underwent end-repair, A-addition and sequencing adaptor ligation, followed by another AMPure bead (Beckman Coulter) clean-up and an enrichment PCR. After a final AMPure clean-up the sample were run on a TapeStation using a High Sensitivity D1000 ScreenTape to confirm the success of the library preparation.

The library was diluted to 1 nM and then further diluted and denatured according to Illumina's recommendations. A final loading concentration of 200 pM, with a 1% spike-in of Illumina PhiX control v.3, was sequenced using paired-end 151 bp reads on a 300 cycle NovaSeq 6000 SP flow cell v.1.5.

2.13 Long-read whole transcriptome sequencing

Nuclease free water was added to 500 ng of globin depleted RNA to create a final volume of 10 µl and added to 1 µl cDNA RT Adaptor and 1 µl Annealing Buffer in a strip tube and was incubated at 60°C for 5 minutes followed by 5 minutes at room temperature. The reagents detailed in Table 2.21 were added to the sample, mixed by flicking the tube and then incubated at room temperature for 10 minutes.

Table 2.21: Components of Nanopore cDNA sequencing ligation mix

Reagent	Volume per reaction (µl)
NEBNext Quick Ligation Reaction Buffer	3.6
T4 DNA Ligase 2M U/ml	1.4
RNaseOUT	1.0

1 µl Lambda Exonuclease and 1 µl USER (Uracil-Specific Excision Reagent) was added to the sample, gently mixed, and incubated at 37°C for 5 minutes. The RNA was purified using 36 µl of RNase-free XP beads, washed twice with 200 µl of Short Fragment Buffer and elution in 12 µl nuclease-free water.

Table 2.22: Components of Nanopore cDNA sequencing reverse transcription mix

Reagent	Volume per reaction (µl)
Maxima H Minus 5x RT Buffer	4.5
Strand Switching Primer II (SSPII)	2.0
RNaseOUT	1.0

1 µl of RT Primer and 1 µl of 10 mM dNTPs were added and incubated at room temperature for 5 minutes, after which the reagents in Table 2.22 were added and the sample incubated at 42°C for 2 minutes. 1 µl of Maxima H Minus Reverse Transcriptase was added to the reaction, which was incubated at 42°C for 30 minutes, followed by 5 minutes at 85°C. The sample was then split into four reactions of 5 µl each and combined with the reagents in Table 2.23.

Table 2.23: Components of Nanopore cDNA sequencing second strand synthesis

Reagent	Volume per reaction (µl)
cDNA Primer	1.5
Nuclease-free water	18.5
2x LongAmp Hot Start Taq Master Mix	25.0

The reactions were placed on a thermal cycler which incubated the reactions at 95^{o}C for 30 seconds followed by 14 cycles of 95^{o}C for 15 seconds, 62^{o}C for 15 seconds and 65^{o}C for 3 minutes and finally an extension of 65^{o}C for 6 minutes. Following PCR ,1 µl of Thermolabile Exonuclease I was added to each reaction and incubated at 37^{o}C for 5 minutes followed by 80^{o}C for 2 minutes. The four reactions were then pooled and purified using 140 µl of AMPure beads followed by two 70 % ethanol washes and elution in 12 µl Elution Buffer. The sample was analysed using both a BR Qubit Assay and a TapeStation with a D5000 ScreenTape to assess its concentration and size profile of the cDNA.

50 fmol of cDNA in 11 µl of Elution Buffer was sequenced as follows. 1.5 µl of Rapid Adapter was diluted with 3.5 µl Adapter Buffer of which 1 µl was added to the cDNA solution and incubated at room temperature for 5 minutes. The PromethION flow cell was prepared and primed as described previously in the section 2.11 Long-read whole genome sequencing. The sample was combined with 37.5 µl Sequencing Buffer and 25.5 µl Library Beads prior to loading onto the flow cell.

2.14 Confirmation of findings by small-amplicon PCR

PCR products were generated for *TTC37* variants using the primers in Table 2.24. The PCR reaction was performed with MegaMix Mastermix (Microzone, UK) comprising of 0.5 µl of sample DNA, 19.3 µl of MegaMix, and 0.1 µl of each 10 µM primer. The reactions were placed on a thermal cycler and incubated under the following conditions: 94°C for 5 minutes, then 30 cycles of 94°C for 30 seconds, 55°C for 1 minute and 72°C for 1 minute, followed by a final 72°C extension for 5 minutes.

Table 2.24: Primers used for confirmation of *TTC37* variants.

Variant	Primer	Sequence	Product size base pairs
c.2808G>A	Forward	*TGTAAAACGACGGCCAGT*AAAATAACTTGTGCTGCTGGAAT	*362*
	Reverse	*CAGGAAACAGCTATGACC*TGATTCTTCTCGTTAGGATTGGA	
c.2634+679A>G	Forward	*TGTAAAACGACGGCCAGT*TGTATGGCAGGAAACAGTCTT	*300*
	Reverse	*CAGGAAACAGCTATGACC*ACTGTTTTGTGTTTCAGCTGGT	

*The italicised portion of the primer is a tagging sequencing allowing use of a universal sequencing primer pair.

PCR products were generated for *TMEM67* variants using the primers in Table 2.25. The PCR reaction was performed with MegaMix Mastermix (Microzone, UK) comprising of 1 µl of sample DNA, 19.3 µl of MegaMix, and 0.1 µl of each 10 µM primer. The reactions were placed on a thermal cycler and incubated under the following conditions: 94°C for 5 minutes, then 30 cycles of 94°C for 30 seconds, annealing for 1 minute (refer to Table 2.24 for annealing conditions) and 72°C for 1 minute, followed by a final 72°C extension for 5 minutes.

Table 2.25: Primers used for confirmation of *TMEM67* variants.

	Annealing Temperature	**Primer**	**Sequence**	**Product size base pairs**
c.1046T>C *TMEM67* exon 10	55°C	Forward	*TGTAAAACGACGGCCAGT*TGATTGGGGCTCTGTGACAT	534
		Reverse	*CAGGAAACAGCTATGACC*CCTCTTGGCTTTGTCTCAGG	
***TMEM67* breakpoint**	60°C	Forward	AGGGTTGGACTTACGATGGT	425
		Reverse	GACAGCATCATTTCACAAACAGT	
***LINC00534* breakpoint**	62°C	Forward	GGTCCAAGAGTCCAAGAGCT	398
		Reverse	CCCCACCTTAAGACCTTCCC	
***TMEM67* multiplex**	60°C	Forward Wt	TGGCTCCTTCATTGACCATG	628 (Wt) + 425
		Forward	AGGGTTGGACTTACGATGGT	
		Reverse	GACAGCATCATTTCACAAACAGT	
***LINC00534* multiplex**	62°C	Forward	GGTCCAAGAGTCCAAGAGCT	398 + 712 (Wt)
		Reverse	CCCCACCTTAAGACCTTCCC	
		Reverse Wt	TACGACAGCCCAAACCAGAC	

*The italicised portion of the primer is a tagging sequencing allowing use of a universal sequencing primer pair. Wt = wild type allele

PCR products were treated with ExoProStar (GE Healthcare Life Sciences, UK) to remove unincorporated primers and dNTPs. 5 µl PCR product was combined with 2 µl ExoProStar and incubated on a thermal cycler at 37°C for 15 minutes followed by 80°C for 15 minutes. Two sequencing reactions consisting of 5.75 µl of nuclease-free water, 1 µl of primer at a concentration of 3.2 pmol/l, 1.75 µl of 5x Sequencing Buffer (Applied Biosystems, USA) and 0.5 µl BigDye v1.1 (Applied Biosystems) were created each containing either the forward or reverse primer. 1 µl of purified PCR product was added to the sequencing reaction in a 96-well PCR plate. The plate was placed on a thermal cycler and incubated under the following conditions: 96°C for 1 minute, followed by 35 cycles of 96°C for 10 seconds, 50°C for 15 seconds and 60°C for 4 minutes. The reactions were purified by ethanol precipitation by adding 1 µl of 3M sodium acetate pH 5.2 and 25 µl of 95% ethanol to each reaction, mixed and incubated at room temperature for 30 minutes. The plate was centrifuged at 2,204 g for 30 minutes, inverted onto paper towel and centrifuged at 180 g for 1 min. 70 µl of 70% ethanol was added to each well and the plate centrifuged at 4°C 1,650 g for 15 minutes. The plate was inverted onto paper towel and centrifuged at 180 g for 1 min to remove residual ethanol before being dried at 95°C for 1 minute. Each reaction was resuspended in 20 µl of Hi Di Formamide (Applied Biosystems) before being sequenced on a 3730 Genetic Analyzer (Applied Biosystems).

3. Towards single cell whole genome sequencing

3.1 Chapter context

At the time this research was initiated, there was a considerable national and international attention on the emerging field of single cell genomics. Consortiums such as the Human Cell Atlas (https://www.humancellatlas.org/), which aimed to map all cell types in the healthy body, were emerging, and their early work defining the single cell transcriptome of the human pancreas had recently been published (Muraro *et al.,* 2016). Locally, we were interested in the possibility of combining our established copy number variation sequencing (CNVseq) workflow (which uses short-read sequencing to identify gains and losses of genomic sequence) with single-cell sequencing. We recognised that this had potential to replace preimplantation genetic screening which was carried out using array-based technology.

3.2 Introduction

The mass of DNA required for NGS library preparation has decreased considerably over the past decade with many ultra-sensitive methods now being commercially available. In recent years much of this work has taken place in the cancer field to enable the detection of tumour-specific mutations. In this setting a much-discussed objective is the implementation of regular screening, post-surgery, to identify resistance to therapy. This scenario relies on so-called liquid biopsies, from minimally invasive peripheral blood samples, to find rare mutations in circulating free DNA from tumours (Pascual *et al.,* 2022). Relevant testing of these samples, in a health-service setting is limited to the targeted identification of specific point mutations in the genome, for which treatment decisions can be made. For example, the UK National Genomic Test Directory (listing the repertoire of funded genomic tests) currently recommends the analysis of ctDNA (when tissue testing is not available) for targeted *EGFR* analysis in patients with locally advanced or metastatic non-small cell lung cancer to determine their eligibility for EGFR inhibitor therapy. Identification of acquired resistance following detection of the EGFR mutation p.(Thr790Met) allows treatment with a third-generation EGFR tyrosine kinase inhibitor osimertinib (Ahn *et al.,* 2019).

Regardless of the progress detecting low frequency point mutations from ctDNA, even now some years after the work presented in this chapter was conceived, the requirement for a nanogram, or more, of starting material for whole genome sequencing still exceeds the of the 6 picograms of genomic DNA present in a single cell. Consequently, whole genome sequencing of germline genomic DNA still necessitates a genome wide amplification step to generate sufficient mass of material for library preparation.

Given the often-limited availability of DNA from the source biological samples and high cost of NGS it is desirable to be able to adjust the required mass and quality of DNA. While commercial whole genome amplification kits were available for the three predominant single-cell WGA chemistries (including degenerate oligonucleotide-primed polymerase chain reaction (DOP-PCR), multiple displacement amplification (MDA) and multiple annealing and looping-based amplification cycles (MALBAC) they remained costly and were sold as proprietary "black box" solutions, with little scope for optimisation. Nevertheless, we reviewed the study by Huang and colleagues which performed WGA of the diploid human BJ cell line (a human foreskin fibroblast), using one of five commercially available kits (each based on one of the available approaches), prior to short-read whole genome sequencing on the HiSeq X Ten platform using 150-bp paired-end reads. Characteristics of the amplified genome (including depth-of-coverage and allele dropout) were dependent on the chosen chemistry. We were encouraged by the high level of sequence coverage, comparable allele dropout rates and lowest false positive rates generated by the multiple displacement amplification WGA kits (Huang *et al.,* 2015a). Based on this report in the contemporary literature we sought to develop an in-house cost-effective custom workflow based on multiple displacement amplification using the readily available reaction components.

3.3 Results

3.3.1 Culturing and verification of control specimens

The experiments reported in this chapter were performed using the cell lines GM12878 and AG16360, which were obtained from the Coriell Institute for Medical Research. They were grown in suspension culture to a confluence of approximately 8 x 10^4 cells/ml. DNA was isolated from the cell pellets, yielding concentrations of 9.74 ng/µl and 10.70 ng/µl for samples GM12878 and AG16360, respectively.

To verify that the cell line AG16360 was trisomic for chromosome 21, the DNA was analysed by microsatellite PCR. Seven markers were amplified and resolved by fragment analysis (Figure 3.1). For two markers, three differently sized alleles were identified, while the remaining 5 were biallelic with a peak height ratio of 2:1, suggesting the sample is aneuploid for chromosome 21.

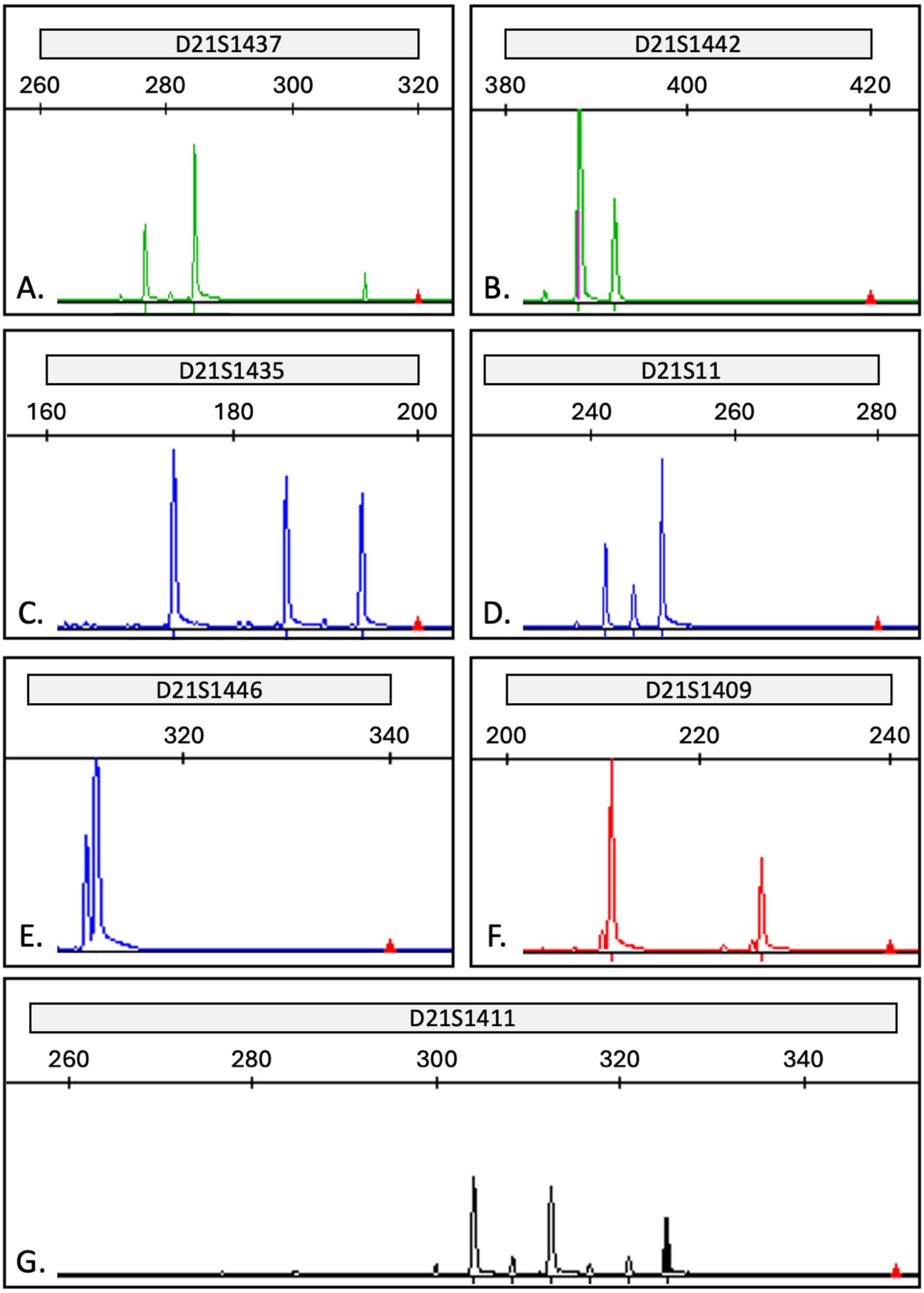

Figure 3.1: Chromosome 21 microsatellite analysis to validate the trisomy 21 cell line. Fragment analysis reveals either a 2:1 (panels **A**, **B**, **D**, **E** and **F**) or 1:1:1 (panels **C** and **G**) ratio of peak heights corresponding to two or three allele sizes respectively. Marker names, corresponding to specific genomic loci on chromosome 21 are recorded in the grey box in each panel. Chromatogram colours correspond to the labelled fluorophore that was used. The horizontal scale bars represent allele sizes in base pairs.

3.3.2 Qubit quantification assay

To develop a robust MDA protocol, WGA was first performed using one nanogram of unamplified genomic DNA. The end-point reactions were quantified using the Qubit system, whereby the fluorescence of a dye is directly related to its binding to double-stranded DNA.

MDA was performed for both samples plus no-enzyme controls for each sample and a no-template control. Post-amplification measurements were performed in triplicate with concentration readings for an exemplar MDA experiment reported in Table 3.1. The absence of double-stranded DNA in the no-enzyme control suggests that amplification was enzyme dependent, but its presence in the no-template control suggests it can be template independent. As more double-stranded DNA was amplified in the no-template controls than in the sample reactions, it is believed that the random hexamers acted as a more efficient template than small quantities of genomic DNA. Consequently, the success of the MDA reaction cannot be determined by the quantification of the double-stranded DNA it created.

Table 3.1: Assessing the creation and concentration of double stranded DNA following multiple displacement amplification.

DNA sample ID	Assay conditions	Quantification values (ng/µl)		
		Replicate 1	Replicate 2	Replicate 3
GM12878	Complete reaction	390	318	344
GM12878	No enzyme control	No reading	No reading	No reading
AG16360	Complete reaction	378	392	350
AG16360	No enzyme control	No reading	No reading	No reading
None	No template control	958	548	520

3.3.3 Assessment of microsatellite genotypes by fluorescent PCR

As MDA amplifies non-template DNA, it was necessary to determine if the amplified DNA could be used as a template in a PCR reaction and, if so, how representative the amplified DNA was of the unamplified sample. The amplified DNA was used as a template in the amplification of a series of microsatellite markers known to be heterozygous in the samples. No chromosome 21 markers were used for sample

AG16360. If a microsatellite marker produced no products or suggested a homozygous genotype, it could be assumed that the amplified DNA did not adequately reflect the sample's genome.

A representative exemplar dataset for this work is presented in Figure 3.2. The unamplified genomic DNA for cell line GM12878 was found to be heterozygous for marker D7S1824 (Figure 3.2: panel A). However, a comparable analysis of DNA amplified from one nanogram of genomic DNA revealed either very small peaks or the loss of an allele (Figure 3.2: panels B, C and D). These findings were similarly inconsistent when using the other markers with AG16360 or GM12878 as a template, with the arbitrary loss of one or both alleles being a common occurrence. In cases where peaks were present but were too faint to analyse, the capillary electrophoresis of the microsatellite marker amplicon was repeated using greater amounts of the amplicon. However, this typically yielded no new insight and was too unwieldy and cumbersome to be used in a routinely in a high-throughput manner. Therefore, it was clear that microsatellite analysis could not serve as a standard quality control tool for MDA products, particularly in a diagnostic context.

Despite these difficulties, microsatellite analysis did consistently show that the no-template control MDA reactions did not generate any amplification products, suggesting that the double-stranded DNA detected by the Qubit assays was not derived from the human genome and so contamination was not an issue. Concatemerization of primer oligonucleotides in the absence of DNA template is a reported artefact of the MDA reaction, and this seemed to be plausible as an explanation for the negative controls in this case.

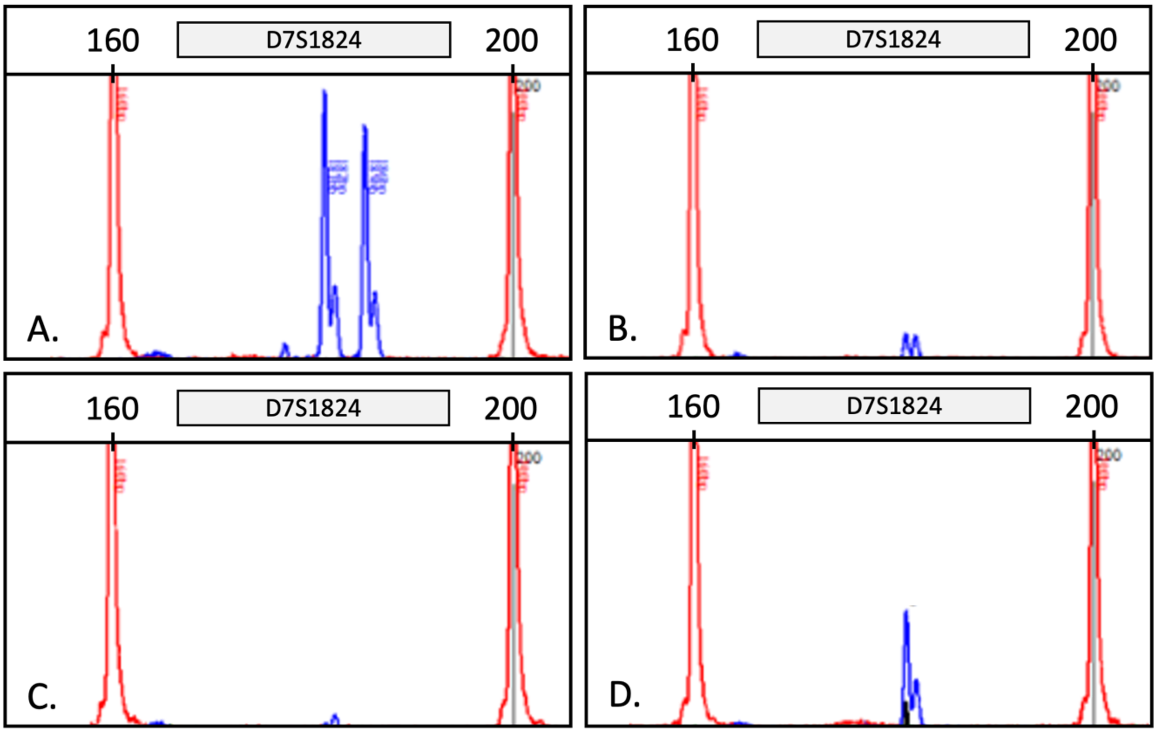

Figure 3.2: A comparison between GM12878 unamplified genomic DNA and MDA reaction amplification products following fragment analysis of the D7S1824 microsatellite. PCRs were performed simultaneously for **(A)** unamplified genomic DNA **(B)** MDA reaction 1 **(C)** MDA reaction 2 and **(D)** MDA reaction 3. Despite all experimental variables being tightly controlled there was considerable variability in assay performance. Apparent allele dropout is observed for MDA reactions, though poor amplification is noted for panel C. The scale bars represent allele sizes in base pairs. Red chromatogram peaks correspond to the reference standard, blue chromatogram peaks correspond to the microsatellite alleles.

3.3.4 Assessment of single nucleotide polymorphism genotypes by TaqMan

To verify the finding of the microsatellite analysis of the MDA products and try and improve on the process of evaluating the success of whole genome amplification, 5 TaqMan SNP assays were used to assess the extent of allele dropout during the amplification process. For GM12878, two of the Taqman target sites were heterozygous while 4 were heterozygous in sample AG16360, although one was on chromosome 21. Each TaqMan assay was performed in triplicate for each MDA product with the results shown in Table 3.2. A blue 'F' indicates a failed assay, which was observed most frequently for sample MDA 1 with 7 out of 15 reactions (47%) failing and all the assays for rs1156253 failing in this sample.

Where a genotype was called, there was no discordance between the expected and experimentally determined genotype for any of the homozygous markers (GM12878: rs6811238, rs747039, rs1156253 and AG16360: rs1109037). However, for heterozygous markers the results were variable with some notable allele dropout occurring. For sample GM12878, all called genotypes for marker rs1109037 were heterozygous as expected. While for the heterozygous marker rs1872575, 4 out 5 successfully amplified reactions were discordant (apparently homozygous) and the fifth genotype was determined to be likely heterozygous although the fluorescent intensities were too dispersed for this to be assigned by the auto caller in the software. It is notable that this marker had the highest number of failed assays (4 out of 9) perhaps suggesting this locus was refractory to amplification by MDA. The four heterozygous markers interrogated for sample AG16360 showed evidence of allele dropout, this was consistent across replicate TaqMan reactions for each MDA-amplified sample. For example, all three replicate assays for marker rs1872575 (MDA 5), rs747039 (MDA 4) and rs1156253 (MDA 4 and MDA 5) had apparently homozygous genotypes (although for the chromosome 21 marker, rs1156253, it was not possible to determine from online data sources which allele was present on the duplicated chromosome 21).

As with the microsatellite analysis, the no-template MDA negative control reactions did not yield a genotype when used as template for a TaqMan assay with any of the markers. As the no-template MDA control reactions amplified DNA as judged by the double stranded DNA specific Qubit assay, it was hypothesised that these samples contained primer-derived product. To assess whether non-template derived amplification products adversely affected the TaqMan assays a titration assay was performed using a combination of genomic DNA and non-template derived MDA products.

NTC (no-template control) MDA-amplified DNA was mixed in a range of ratio (0:1, 1:4, 2:3, 3:2 and 4:1) with unamplified genomic DNA from sample GM12878 to create a titration series on which the genotype of rs1872575 was determined using the TaqMan assay (Figures 3.3A to E). This demonstrated the presence of NTC material had no impact on the reproducibility of heterozygous calls; however, there was a minimal reduction in signal intensity of the assays the proportion of NTC increase. As this didn't affect the accuracy of the assay, this was interpreted to be due to a reduction in the level of target template rather than an inhibition of the assay by the NTC product. Consequently, the results of this titration assay support the view that discordant microsatellite and TaqMan results were due to uneven allele/loci amplification by the MDA reaction rather than inhibition of the reactions by non-template derived amplification product.

Table 3.2: TaqMan resolved genotypes for MDA reactions generated from one nanogram of DNA extracted from cell lines GM12878 and AG16360.

dbSNP identifier	Chr		GM12878											AG16360									
			Ref	MDA 1			MDA 2			MDA 3				Ref	MDA 4			MDA 5			MDA 6		
				R1	R2	R3	R1	R2	R3	R1	R2	R3			R1	R2	R3	R1	R2	R3	R1	R2	R3
rs1109037	2		+/-	+/-	+/-	F	F	+/-	+/-	+/-	+/-	F		+/+	+/+	+/+	+/+	F	+/+	+/+	+/+	+/+	+/+
rs1872575	3		+/-	+/+	F	F	+/+	F	+/+	+/+	?	F		+/-	?	+/-	?	+/+	+/+	+/+	?	+/+	?
rs6811238	4		+/+	F	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+		+/-	F	?	?	+/-	+/-	+/-	?	+/+	+/+
rs747039	17		+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+		+/-	+/+	+/+	+/+	+/-	+/-	+/-	?	+/-	+/-
rs1156253	21		+/+	F	F	F	+/+	+/+	+/+	+/+	+/+	+/+		+/-	+/+	+/+	+/+	+/+	+/+	+/+	?	?	?

Chr: Chromosome. **R1:** Replicate 1. **R2:** Replicate 2. **R3:** Replicate 3. **Ref:** Reference genotype. **+/+:** Homozygous genotype. **+/-:** Heterozygous genotype. **F:** Failed TaqMan amplification reaction. Red genotypes are those that are discordant between the known reference genotype and the experimentally determined genotype generated using the sample from the MDA-reaction. **?:** indicates a likely heterozygous genotype (the datapoints on the X/Y scatter plots of per-allele fluorescence were not consistent with a homozygous reference genotype, or homozygous non-reference genotype, and were insufficiently tightly clustered for the analysis software to assign a genotype).

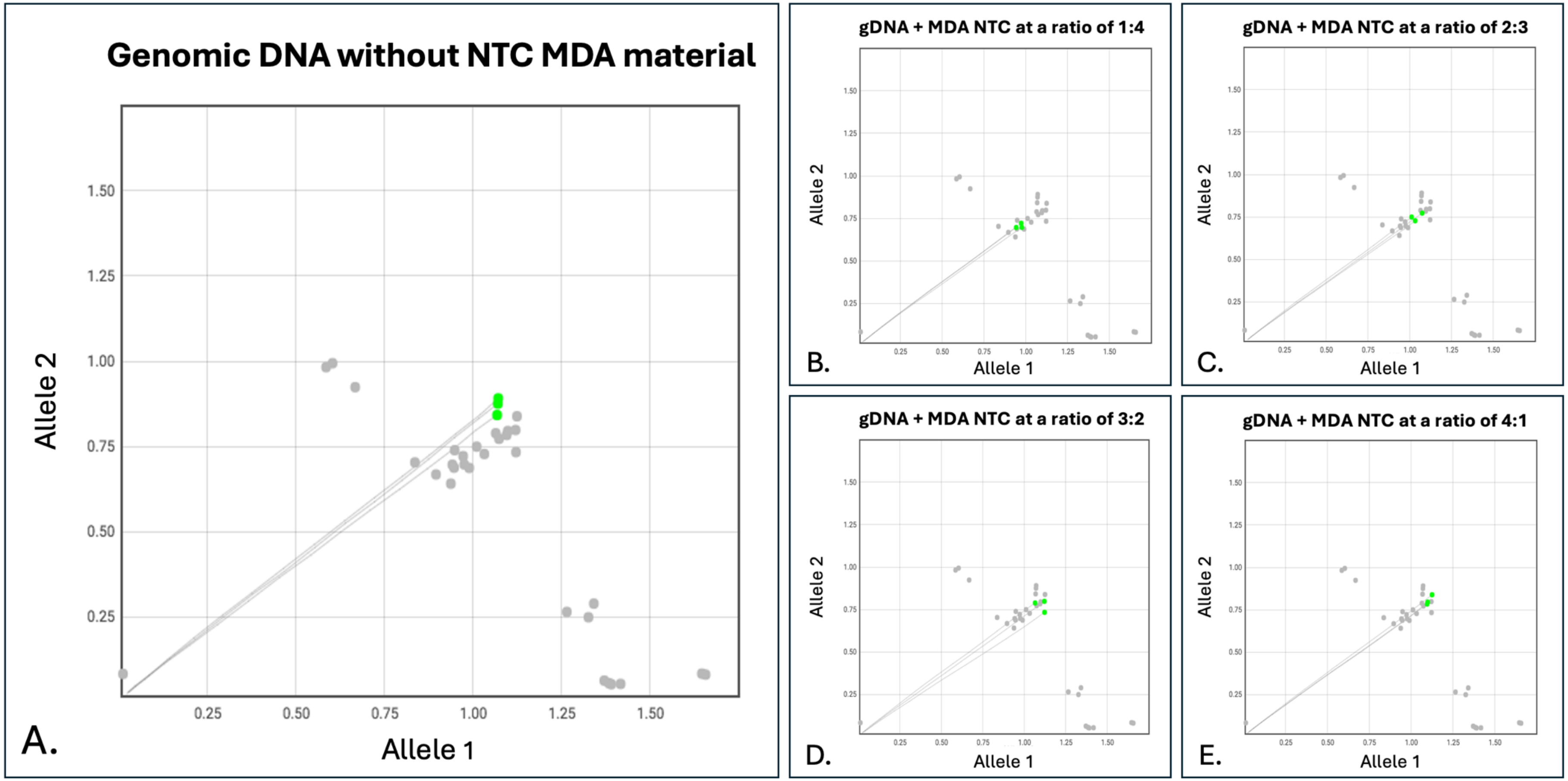

Figure 3.3: TaqMan genotyping plots demonstrating the apparent concatemerized primer product detected in the negative controls does not impact genotyping of genomic DNA. (A) Unamplified genomic DNA for sample GM12878; heterozygous genotypes are highlighted green and are tightly clustered. NTC MDA material was added to GM12878 unamplified genomic DNA at increasing relative ratios (B) 1:4 (C) 2:3 (D) 3:2 (E) 4:1. The addition of NTC MDA material did not affect the clustering of heterozygous genotypes, indicating the double stranded products did not interfere with the assay. Fluorescence of each allele is measured on the X (allele 1) and Y axis (allele 2) respectively.

3.3.5 Characterisation of MDA samples by next generation sequencing

The genotyping assays performed on the MDA material thus far demonstrated there had been variability in yield and allele dropout for interrogated microsatellite and single nucleotide markers. To better characterise the composition of MDA-amplified sequences, whole-genome next-generation sequencing was performed on the samples. The NTC MDA reactions were also sequenced, in order to better understand the composition of the material generated in these "blank" reactions.

Library preparation was successful for the unamplified genomic DNA, all three of the NTC MDA reactions and five of the six MDA reactions - library production for one of the AG16360 MDA replicates, MDA4, failed. The libraries were pooled and sequenced on a NextSeq 550 to create 151 bp paired end reads. The average cluster density over four lanes was 190 k/mm^2 with 92.7% of clusters passing the sequencer filtering criteria. The sequencing yielded 91.63 gigabases of data, with 81.4% of bases achieving quality scores equal to or greater than Q30; this is consistent with the manufacturer's performance expectations, indicating that the MDA material could be prepared for sequencing.

Sequence reads were aligned to the human reference genome GRCh37, this earlier genome build having been utilised to develop the CNVseq bioinformatics pipeline we were aiming to use for this workflow. The per-chromosome alignment count was noted, both by comparing to the alignment count observed for the unamplified genomic DNA samples and the expected read count based on the length of each chromosome (Figure 3.4). Unsurprisingly, the expected per-chromosome read distribution and the distribution for the unamplified genomic DNA samples were most closely correlated to each other. While the overall trend for the MDA material demonstrated fewer reads mapped to smaller chromosomes than expected, the variability between samples was considerable. For some chromosomes, there was significant deviation from the expected value. For example, for sample MDA 1 (GM12878 Figure 3.4A), the expected proportion of reads mapping to chromosome 12 was 4.65%, but the observed value was 7.66%. Similarly for sample MDA 5 (AG16360

Figure 3.4B), the expected proportion of reads mapping to chromosome 20 was 2.19%, but the observed value was 5.36%.

Given the variability in the per-chromosome distribution of sequence reads between MDA samples, specific loci from each autosomal chromosome arms were visualised using the Integrative Genomics Viewer (Robinson *et al.,* 2011). Each locus demonstrated similar characteristics, and so the extensively studied region containing the hereditary breast cancer gene *BRCA1* will be used as an exemplar and discussed in greater detail. Results for the samples GM12878 and AG16360 are displayed in Figures 3.5 and 3.6, respectively. For both unamplified genomic DNA samples, the aligned reads show the expected uniform distribution across the region. By contrast, the distribution of alignments across the region for each of the MDA reactions was very uneven. While alignments were present across the regions, there were areas with little to no cover, punctuated with intervals with very high read counts. The location of these areas of minima and maxima read counts appeared random and showed no consistent pattern across samples or with underlying genomic features. These data suggest the initial MDA prime and extension events of an MDA reaction are critical to achieving robust whole genome amplification.

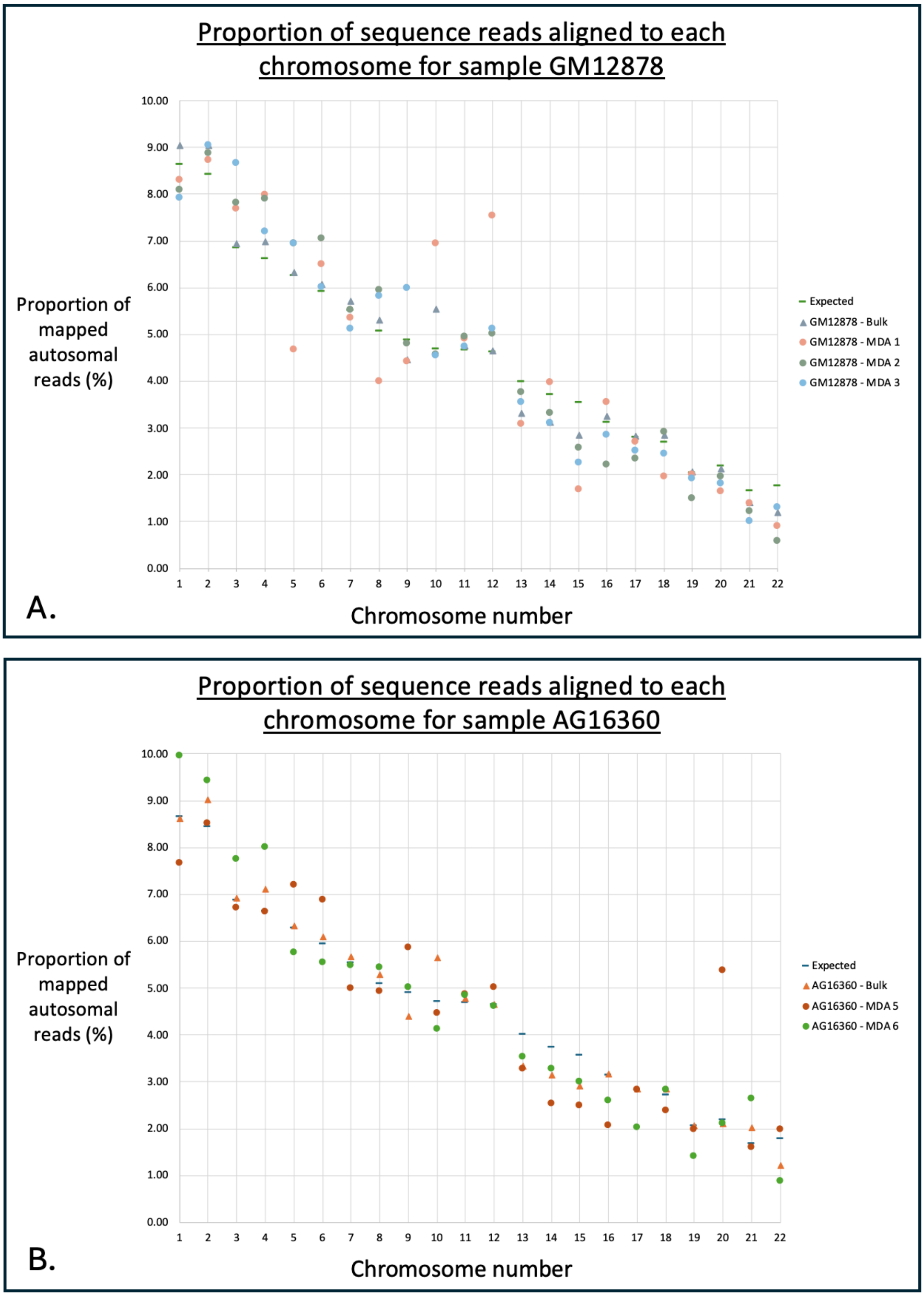

Figure 3.4: The per-chromosome distribution of aligned sequence reads generated from gDNA and MDA amplified material. While the trend of fewer sequence reads mapping to smaller chromosomes is consistent for both **(A)** GM12878 and **(B)** AG16360 samples, there is considerable variation between different MDA generated material. The proportion of reads for the unamplified genomic DNA is typically closest to the expected value. The expected proportion of reads is based on chromosome length.

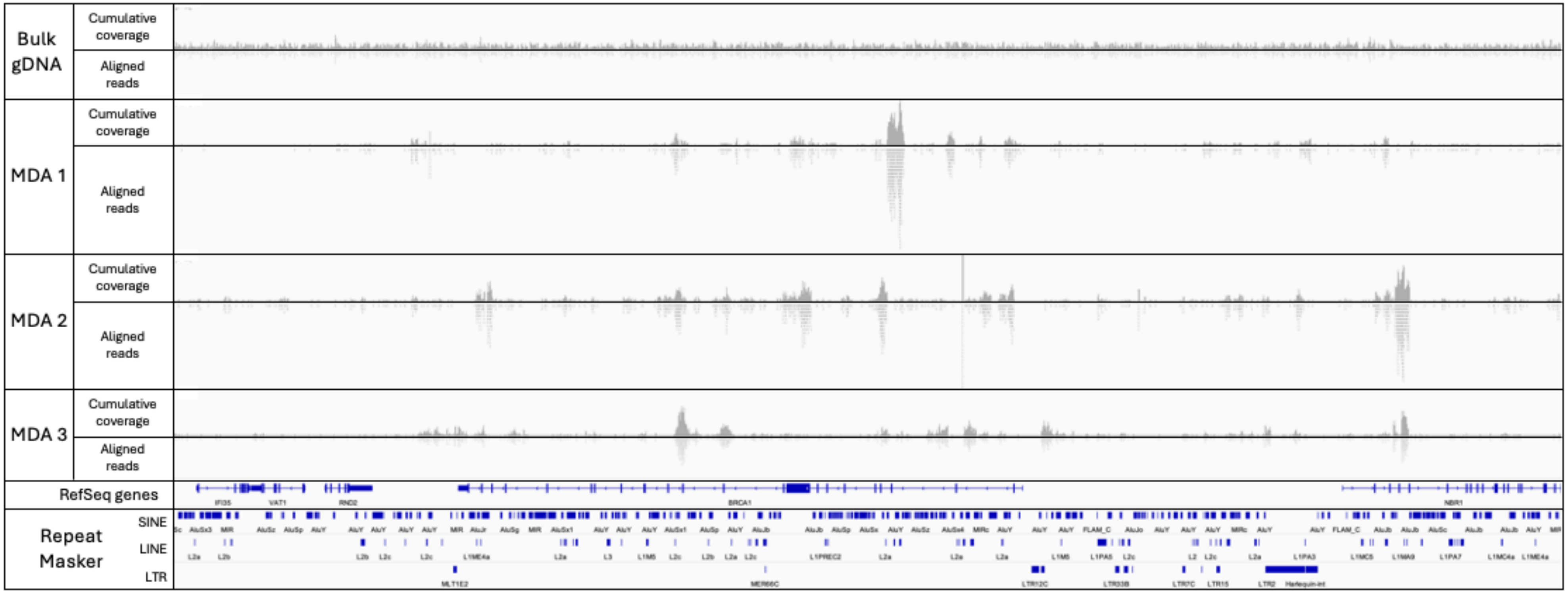

Figure 3.5: IGV rendered alignments of MDA material generated from sample GM12878. The unamplified genomic DNA sample shows an even read distribution across the displayed *BRCA1* locus. Contrastingly the MDA products show regions of high coverage that are poorly replicated between MDA reactions. The *y*-axis of cumulative coverage tracks is scaled 0-30×.

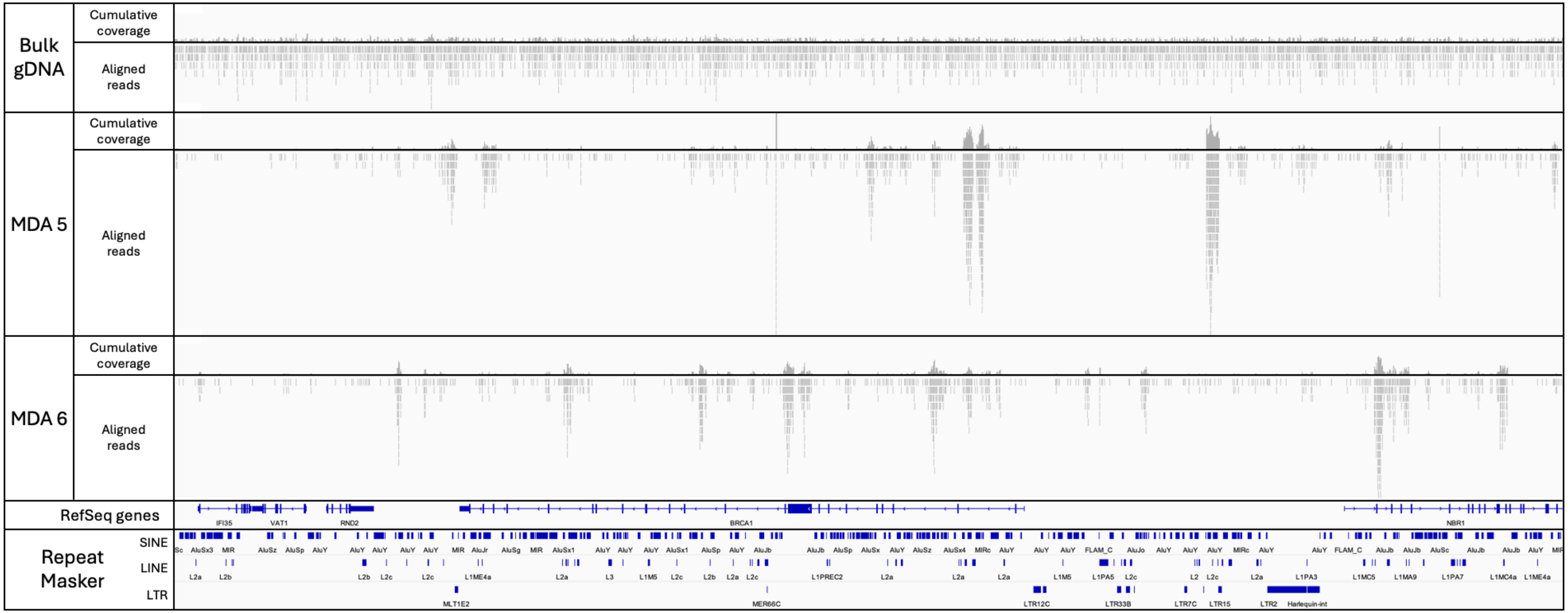

Figure 3.6: IGV rendered alignments of MDA material generated from sample AG16360. The unamplified genomic DNA sample shows an even read distribution across the displayed *BRCA1* locus. Contrastingly the MDA products show regions of high coverage that are poorly replicated between MDA reactions. The *y*-axis of cumulative coverage tracks is scaled 0-30×.

When visualised using a TapeStation, the fragment size profile of libraries generated from the NTC MDA products was comparable to those derived from unamplified genomic DNA. Nevertheless, despite all the libraries being pooled in equimolar concentrations, the per-library read count for the NTC MDA libraries was markedly lower than the other libraries (Table 3.3). For these samples, approximately half the total number of reads were generated, suggesting they were less able to form clusters on the flow cell compared to the other libraries. Furthermore, the PCR duplicate rate was substantially higher across all NTC MDA samples (35.28% to 45.12%) compared to the other libraries. In comparison, the unamplified genomic libraries had PCR duplicate rates below 3%. High PCR duplicate rates suggest the NTC MDA libraries contained very few unique sequences when compared to unamplified genomic DNA libraries. When the proportion of reads that could be aligned to the human genome was determined, surprisingly, up to 51% of sequence reads in the NTC MDA sample libraries were aligned. However, this is considerably less than the >98% of reads from the unamplified genomic DNA or MDA reaction libraries.

Since the aligned sequences in the NTC MDA libraries could represent contamination and impact the value of any analysis preformed on amplified DNA regions, the alignments were visually inspected. Regions of high (>500×) sequence depth were identified and manually reviewed using the IGV browser; an exemplar locus is reported in Figure 3.7. It was seen that aligned reads were typically mapped to low-complexity sequences as well as SINE and LINE repeats as indicated by the RepeatMasker tract (Robinson *et al.,* 2011). This suggest that the sequences derived from the NTC MDA samples did not originate from a human genome but were ‘random’ sequences that were able to weakly align to low complexity reads in the human genome. As was observed for the MDA reactions from unamplified genomic DNA the locations of high sequence coverage in the NTC MDA library alignments were not reproducible between NTC MDA samples.

To investigate whether short-read NGS data could provide insight into the discordant TaqMan genotyping results the genomic position of each marker was reviewed. Insufficient read depth at the interrogated loci meant that it was not possible to determine whether there was no or partial (one allele) whole genome amplification at these sites for each of the MDA reactions.

Table 3.3: Summary of sequencing and alignment metrics.

Source specimen	MDA number	Total reads		PCR duplicate reads		Reads aligned to reference genome	
		Total count	Proportion of pool (%)	Total count	Proportion of library (%)	Total count	Proportion of library (%)
GM12878	N/A (Bulk)	105,408,595	10.98	2,402,586	2.28	104,338,937	98.99
GM12878	MDA 1	99,614,341	10.38	16,412,048	16.48	98,831,257	99.21
GM12878	MDA 2	120,022,527	12.50	5,307,039	4.42	119,653,416	99.69
GM12878	MDA 3	116,711,388	12.16	10,740,918	9.20	116,112,194	99.49
AG16360	N/A (Bulk)	104,052,110	10.84	25,02,985	2.41	103,813,975	99.77
AG16360	MDA 5	115,354,907	12.02	12,273,426	10.64	114,796,013	99.52
AG16360	MDA 6	120,786,889	12.58	9,615,784	7.96	120,316,581	99.61
NTC	MDA 13	45,554,812	4.75	20,556,236	45.12	23,336,783	51.23
NTC	MDA 14	56,645,850	5.90	22,890,588	40.41	26,895,450	47.48
NTC	MDA 15	75,697,337	7.89	26,702,945	35.28	35,113,080	46.39

NTC: No template control. **N/A:** Not applicable (corresponds to unamplified genomic DNA without MDA amplification).

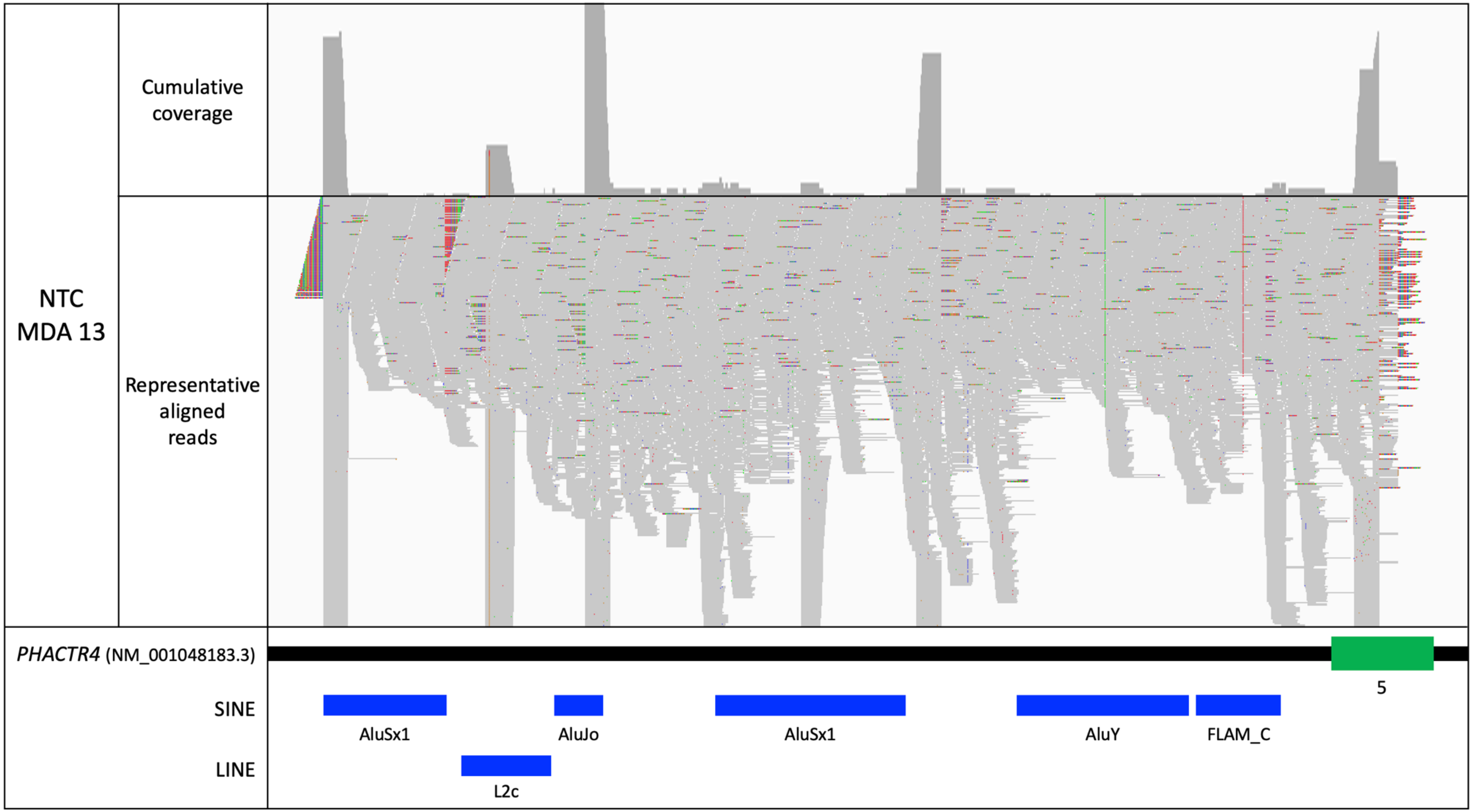

Figure 3.7: IGV rendered alignments at the representative *PHACTR4* gene region for the NTC MDA 14 reaction. Sequences intersect repetitive SINE and LINE elements defined by the RepeatMasker tract. The NTC MDA reactions MDA14 and MDA15 were also reviewed but had no reads mapping to the locus (data not shown). The *y*-axis of cumulative coverage tracks is scaled 0-3,300×. The displayed interval is hg19.chr1:28,790,064-28,792,848. NTC: No template control.

3.3.6 Detection of copy number variation from NGS data

Despite read distribution being very variable across each chromosome (Figure 3.4), we sought to determine whether CNVseq (Section 2.8.2), a previously validated NGS-based in-house copy number variation pipeline, could be deployed for use with MDA-generated sequencing libraries. Briefly, aligned sequence reads are counted across a sliding window of genomic intervals with the value compared to the comparable value derived from a reference dataset comprising a pool of normal controls. A representative karyogram, created by CNVSeq, for chromosome 1 is displayed in Figure 3.8. While the sequencing library generated from the GM12878's unamplified genomic DNA (Figure 3.8A) displayed a normal copy number profile (comparatively tightly clustered windows). However, the read counts for the sliding window for MDA samples (MDA 1 in Figure 3.8B) were very variable and produced a copy number value (red line) that was below the normal baseline (i.e., a relative copy number of 1.0). This was seen across all chromosomes in the analysed samples (MDA5 and MDA6).

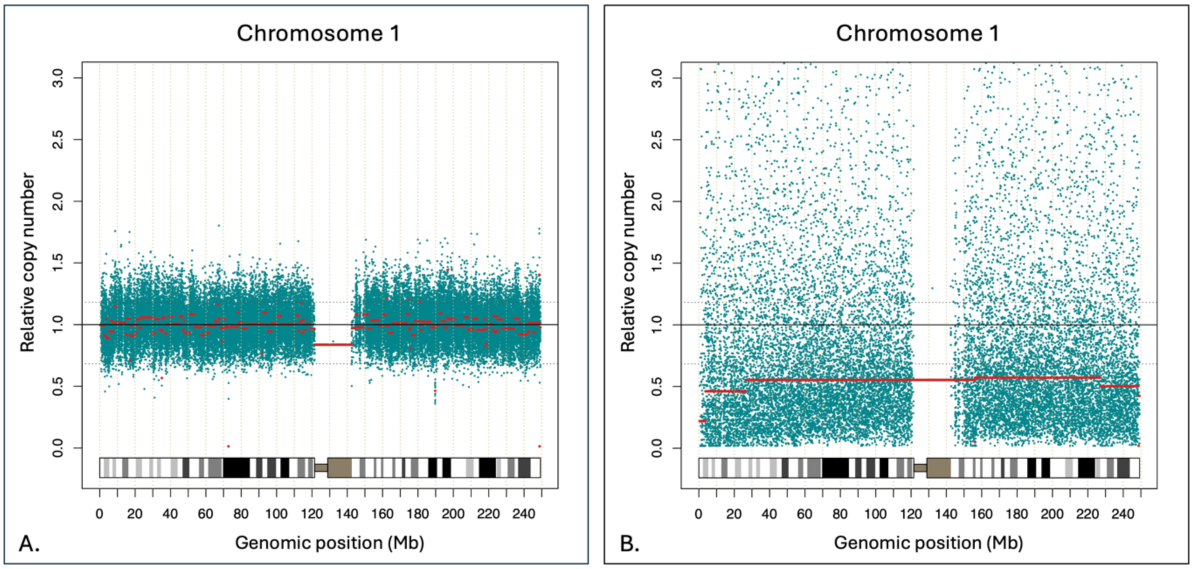

Figure 3.8: Chromosome 1 CNVseq karyograms generated from sliding window counts of aligned sequence reads. The relative copy number for each window (turquoise spot) is plotted, and when consecutive windows deviate from the normal relative copy number (1.0), the interval is segmented with a red bar. **(A)** A normal relative copy number profile for unamplified genomic DNA from sample GM12878.Note the homozygous deletions defined by the red dots showing 0.0 relative copy number at positions chr1:72,766,693-72,804,769 and chr1:248,734,098-248,794,371. **(B)** An exemplar profile generated from sample MDA 1. The dispersed relative copy number of these windows and segmentation baseline that is below 1.0 prohibited analysis of these data. Note the lack windows at the centromere

where repetitive sequence prohibits unambiguous alignment of sequence reads. Genomic positions are reported according to human genome build hg19.

CNVseq data processing for AG16360's unamplified genomic DNA sample revealed the expected segmentation profile, suggesting a relative copy number of 1.5 (Figure 3.9A). Like the karyogram for chromosome 1 (Figure 3.8b), the read count values of the sliding windows for sample MDA 5 were very variable, and the predicted copy number was not increased to the expected relative copy number of 1.5 (Figure 3.9B).

Increasing the length of the sliding window caused the window's read count to increase. To determine whether this would lead to CNVSeq predicting a copy number of 1.5× for chromosome 21, window sizes of 160, 500, 1000 and 2000 bp were generated (Figure 3.9C-F). Increasing the window size did improve the viability of the read counts for each window. However, while increasing the window size to 500 bp did result in an increase in predicted copy number, it didn't reach the expected value of 1.5x. Subsequent increases in window size had a negligible effect on the reported copy number of the majority of chromosome 21, while also predicting the last 10 Mb of chromosome 21's p arm had a copy number of approximately 2x.

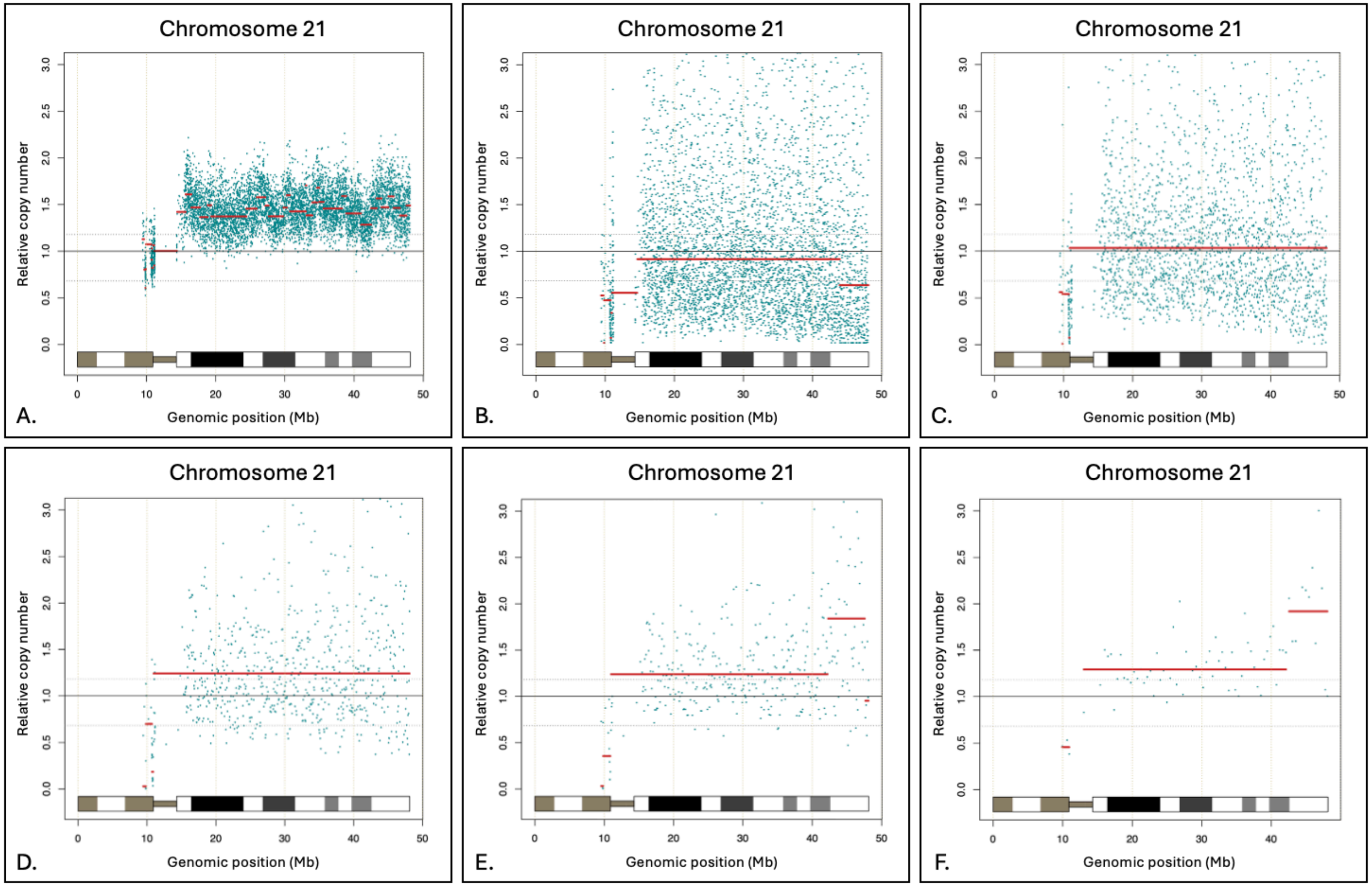

Figure 3.9: Chromosome 21 CNVseq karyograms for aneuploid sample AG16360. (A) The unamplified genomic DNA sample is trisomic for chromosome 21. Note the tight distribution of windows and segmentation line showing an increase in the relative copy number to 1.5. **(B)** Sample MDA 5 run through the pipeline revealing dispersed windows and no detectable shift in segmentation. To adjust the pipeline the size of the genomic window was increased to **(C)** 160 bp, **(D)** 500 bp **(E)** 1000 bp and **(F)** 2000 bp. This had no effect on the ability to detect the aneuploidy. Genomic positions are reported with respect to human genome build hg19.

3.4 Discussion

The presented data sought to assess the suitability of multiple displacement amplification as a precise and accurate method for whole genome amplification. This work was motivated by the ambition to develop this methodology into a single-cell workflow for preimplantation genetic screening of embryo biopsies.

Investigations were carried out using two comprehensively characterised lymphoblastoid cell lines, which can be used as a long-term source of high molecular weight DNA. Sample GM12878 had been sequenced using a range of sequencing technologies as part of the Genome in a Bottle Consortium's endeavour to create a set of high-confidence genotypes that can be used by researchers to benchmark workflows in their own laboratory. Similarly, sample AG16360 is a widely available cell line that is trisomic for chromosome 21. As whole chromosome aneuploidy is a major cause of pregnancy loss, experiments using this sample were consistent with the eventual aim of developing the workflow for the analysis of embryo biopsies for PGS.

To develop expertise in the amplification and quality assurance of MDA reaction products, unamplified genomic DNA (rather than single cells) was used from the outset. The MDA reactions were developed with a one nanogram of DNA, the equivalent of approximately 160 human cells. The observation that, by Qubit quantification, amplification had only occurred in presence of the Phi29 polymerase provided confidence that the enzyme had catalysed the process. Nevertheless, the identification of a quantifiable product in the no-template control reaction provided some cause for concern. For most molecular biology workflows, including those related to next-generation sequencing, there is typically little interest in the characterisation of no-template control reactions. However, the no-template reaction products were retained and could therefore be further characterised alongside those that were generated from the GM12878 and AG16360 samples.

The selection of MDA (rather than DOP-PCR or MALBAC) as the initial approach to WGA was based on published metrics showing a high level of sequence coverage, comparable allele dropout rates and the lowest false positive rates across several commercially available kits (Huang *et al.,* 2015b). The allele dropout rate was investigated to improve the characterisation of MDA-WGA reactions beyond the binary metric of whether or not amplified DNA is present. Allele dropout was assessed with the use of repurposed microsatellite markers that had been initially developed for genome-wide linkage studies, with allele dropout determined when a known heterozygous marker was found to be homozygous. Consequently, I selected markers known to be heterozygous in the GM12878 and AG16360 samples using publicly available genotype information. It was apparent that while the fluorescently labelled PCR reagents generated strong and robust chromatograms for unamplified genomic DNA samples, their use with MDA reactions was less straightforward. Despite maintaining tightly controlled reaction variables (*e.g.* the use of a single PCR master mix, concurrent amplification on the same thermocycler, and the same mass of input DNA), their amplification profile was highly variable. Differences in peak heights between alleles suggested the underlying allele fraction was inconsistent with the expected 50:50 allelic ratio.

Loading a serial dilution, as well as the undiluted microsatellite amplification products from these samples, onto a 16 capillary 3130xl Applied Biosystems capillary electrophoresis instrument became time-consuming and unwieldy. This, coupled with the need to perform a preliminary run to determine the appropriate dilution factor, limited the utility of using microsatellites to determine whether the sample had been amplified and whether allele dropout had occurred. Nevertheless, the persistent lack of amplification for the NTC MDA reactions, for any of the microsatellite markers tested, suggested that the concentration readings detected by Qubit assays were unlikely to represent template contamination from neighbouring reaction wells but instead were probably double-stranded DNA derived from the reaction's oligonucleotides self-priming.

The TaqMan genotyping assay was subsequently identified as a more suitable assay to determine allele dropout, as it is potentially sensitive and easier to scale to routine diagnostic use. Focus therefore shifted from the analysis of microsatellite markers to single nucleotide polymorphisms. Following the strategy used for microsatellite genotyping, markers were selected based on their known heterozygous genotypes in the GM12878 and AG16360 cell lines. For unamplified genomic DNA samples, the 'autocalling' function of the TaqMan software could robustly identify heterozygous genotypes from each allele's level of fluorescence. Despite this success, TaqMan reactions for MDA amplified samples frequently remained undetermined.

Manual inspection of allele-specific fluorescence data frequently supported the presence of both alleles for known heterozygous positions, allowing the genotypes to be manually determined. Nevertheless, there remained a considerable number of discordant genotypes. Where the TaqMan assay repeatedly failed to produce a signal for either allele, it was assumed that the locus had not been amplified in the MDA reaction. When a signal was produced, it was noted that two patterns of discordant genotypes occurred by the TaqMan assays: all the genotypes for an MDA sample were the same but wrong, or the assays gave a series of apparent random calls. It was hypothesized that the former occurred when the locus was amplified to a level below the TaqMan's limit of detection, while the latter was due to the preferential amplification of one allele.

Based on these three causes of genotyping error, it was determined that the instances of allele dropout were observed to affect 1/3 of loci in the amplified samples. For GM12878, 2 out of 6 loci failed: rs1872575 MDA 1 and MDA 2, while for AG16360, 4 out 12 loci failed: rs1872575 MDA 5, rs747039 MDA4, rs1156253 MDA4 and MDA5 (Table 3.2).

An alternative approach to TaqMan genotyping would have been to use a medium-throughput platform such as the MassARRAY® system (Agena Bioscience). This system enables the simultaneous analysis of up to 40 single nucleotide markers and can be deployed in a 384-well format. Molecular diagnostic laboratories have frequently adopted medium-throughput genotyping assays to create a "molecular fingerprint" for samples, which can be used in combination with other laboratory tests to alert analysts to potential sample mix-ups or DNA contamination. A standard set of single nucleotide markers has been proposed that are sufficiently polymorphic in the general population to generate a combination of heterozygous genotypes that can characterise an individual (Pengelly *et al.,* 2013). However, no suitable medium-throughput instrument was available for this study.

To assess the uniformity of MDA-WGA, samples were characterised by low-coverage WGS. Sequencing libraries were successfully generated for five of six MDA samples. Despite sample MDA 4 producing a comparable quantity of DNA whose microsatellite PCR and TaqMan analysis yielded comparable results to the other WGA reactions, for unknown reasons it failed to produce an NGS library.

Randomly selected p- and q-arm loci, for each chromosome, were manually reviewed, revealing a non-uniform distribution of sequence reads for MDA samples, compared to libraries generated from the respective unamplified genomic DNA. The genomic intervals under investigation either showed no mapped sequence reads or reads with varying densities. A complete absence of mapped reads across a large region suggested that the MDA reaction failed due to an underlying quality of the sequences, such as their GC content or level of sequence complexity. By contrast, where read coverage was variable, it was often not clear whether the amplification favoured one allele over the other.

To further investigate the allelic ratio of amplified sequences, the sequencing depth at the chromosome positions that were interrogated by microsatellite and TaqMan genotyping was manually reviewed. For all loci that were inspected, the depth of coverage was too low to be informative, with most positions linked to no reads. While this could have been resolved by resequencing the samples at greater depth, this was not done due to the prohibitive cost.

Despite the non-uniformity of sequence read distribution, NGS data was cautiously processed using an in-house copy number variant sequencing (CNVseq) pipeline. The workflow uses a sliding window to count the number of reads across a genomic region and compares this value to the comparable number of mapped reads from a control library made from the DNA from 10 “normal” males. The CNVseq workflow had previously been validated for the identification of copy number gains and losses, including whole chromosome aneuploidy, by the local diagnostic genomics laboratory. The read count for each window’s relative copy number displayed a tight distribution for libraries prepared from unamplified genomic DNA, and for sample AG16360 a gain of chromosome 21 was correctly identified. Contrastingly, the relative copy number of windows for the MDA reactions was highly disparate and highly variable; this was interpreted to have been caused by the non-uniform WGA. Increasing the window size reduced the level of variation, while also reducing the test resolution. Nevertheless, it was still not possible to resolve chromosome 21’s copy number for the MDA 5 and MDA 6 reactions.

MDA uses the highly accurate Phi29 DNA polymerase for non-specific amplification with minimal error, with the new strand displacing the original strand, allowing for exponential amplification without the requirement for thermally induced denaturation steps. Early comparative studies on whole genome amplification supported the use of MDA compared to other PCR-based methods for whole genome analysis (Pinard *et al.,* 2006), including directly from clinical samples (Hosono *et al.,* 2003). Nevertheless, when used with very small amounts of input template, poor repeatability and allele loss has been reported (Sidore *et al.,* 2016). To overcome these limitations, digital droplet MDA (ddMDA) was proposed. In this modified method, compartmentalisation of single molecules of template DNA into picolitre-sized droplets enables the amplification of single template molecules to completion (Sidore *et al.,* 2016). However, this methodology remains challenging to implement due to the need to use a commercial instrument for microfluidic emulsification. Furthermore, it is difficult to obtain a reliable source of the chemically inert oils and surfactants that are required to generate the droplets. The data presented in this study was generated using MDA products amplified from one nanogram of input DNA, which is approximately equivalent to the DNA in 166 human cells. Had the mass of the template been lowered to single-cell levels, it is likely that the number of discordant genotypes would have increased, making the analysis of little diagnostic value.

While the presence of double-stranded molecules in the NTC MDA reactions was believed to be due to self-primed products, it is notable that only 50% of reads mapped to the human reference genome. Ad-hoc manual inspection of these sequences indicated they were mapped to low-complexity repetitive sequences. The source of the remaining sequences was unknown but could have arisen from low-level bacterial contamination, either of the reagents or laboratory atmosphere (Glassing *et al.,* 2016). Reagent contamination has been considered by other investigators, and methods to create a “high purity” MDA reagent that includes an assay to directly measure contaminant levels have been previously reported (Blainey & Quake, 2011). More straightforwardly, investigators have reported the effective use of UV irradiation to degrade contaminating DNA in the suppression of the amplification of unwanted target DNA (Woyke *et al.,* 2011).

The continued development of sequencing instruments, including the release of third generation long-read instruments, presents an opportunity to sequence long DNA fragments. The use of isothermal amplification, with methodologies such as MDA, has the benefit of generating fragments longer than 10 kb (and up to 100 kb), which are potentially suitable for analysis by long-read sequencing. Nevertheless, the reported formation of chimeric DNA molecules where the 3’ end of a displaced strand anneals and then extends along an unrelated sequence highlights the challenge when considering the clinical application of long-read sequencing of WGA products (Lasken & Stockwell, 2007) establishing a phased parental haplotype in the presence of chimeric DNA molecules is unlikely to be successful.

The motivation to commence this work was prompted by emerging interest in the single-cell characterisation of embryo biopsies by whole genome sequencing. Since this time comprehensive guidance has been issued by the European Society of Human Reproduction and Embryology (Coonen *et al.,* 2020) both with respect to the approaches used to obtain the biopsied sample and the suitability of genomic technologies, such as analyses by fluorescence *in situ* hybridisation, aCGH and NGS. While the clinical use of PGS in the embryology setting remains intensely debated, its use is suggested if the following indications have been reported: advanced maternal age (greater than 35 years), recurrent implantation failure (three or more failed IVF-embryo transfer cycles involving high-quality embryos), major male fertility contraindications (a sperm sample with a combination of oligospermia, asthenzoospermia and/or teratozoospermia or azoospermia) (Carvalho *et al.,* 2020). By contrast, couples with a history of recurrent miscarriage have a high chance of successfully conceiving, and PGS in patients without a known genetic issue is not recommended (ESHRE Guideline Group on RPL *et al.,* 2018).

3.5 Summary

Assays used to characterise the performance of MDA-WGA reactions highlighted several challenges with the chosen approach. Biallelic amplification was often biased and incomplete, while sequence coverage was highly variable meaning many positions could not be interrogated with confidence.

This work was largely completed prior to a first period of maternity leave; during this time a comprehensive restructuring of England's NHS regional genetic diagnostic laboratories was announced. An online test directory (which specifies the genomic tests commissioned by the NHS) was under development (https://www.england.nhs.uk/publication/national-genomic-test-directories/), and it was clear that single-cell derived PGS-based analyses would not be an emerging priority. This, in combination with the difficulties experienced with the presented experiments, prompted a change in direction to the analysis of hereditary cancer, the subject of the next chapter of this thesis.

4. Investigating linked-read target enrichment to assess pseudogene sequences at the *PMS2* locus

4.1 Introduction

While short-read next generation sequencing has revolutionised the speed and number of genomic loci that can be concurrently analysed in a single assay, some regions remain difficult to sequence. Typically, this is either due to the region being difficult to amplify using the available reagents (*e.g.* due to a high GC-content), or because the resulting sequence reads cannot be unambiguously aligned to the reference genome. Genome wide analyses have identified genes that are medically relevant and may require additional consideration when designing and interpreting clinical assays (Mandelker *et al.,* 2016). Interference by non-functional pseudogenes, located elsewhere in the genome, is a typical cause of sequence reads being inappropriately mapped and thereby reducing the sensitivity for variant detection.

To overcome the limitation of ambiguous mapping several commercial solutions have been developed that generate synthetic "long-reads". By introducing a molecular index into the original "long" DNA fragment the genomic origin of the sequence can be assigned. One such method was developed by 10X genomics and is termed linked-read sequencing (Zheng *et al.,* 2016). While this was developed to be a whole-genome approach to synthetic long-read sequencing we sought to use the library preparation method, in combination with a custom-designed hybridisation capture reagent, to perform targeted sequencing of the pseudogene-containing *PMS2* locus located on chromosome 7p.

Analyses of *PMS2*, by short-read sequencing, are confounded by the presence of the *PMS2CL* pseudogene; the homologous region contains *PMS2* exons 9-15 (but excludes *PMS2* exon 10). By generating linked-read molecules and selecting only those that are anchored within the gene-specific region, it was anticipated that the specificity of alignments could be improved. To identify sequence reads mapping to the 3′ end of exon 15, fragments that are at least 24.7 kb in length were required. The interrogation of other *PMS2* pseudogene affected regions would require shorter assembled molecules; an assessment of the minimum and maximum length of linked-read molecules is displayed in Figure 4.1.

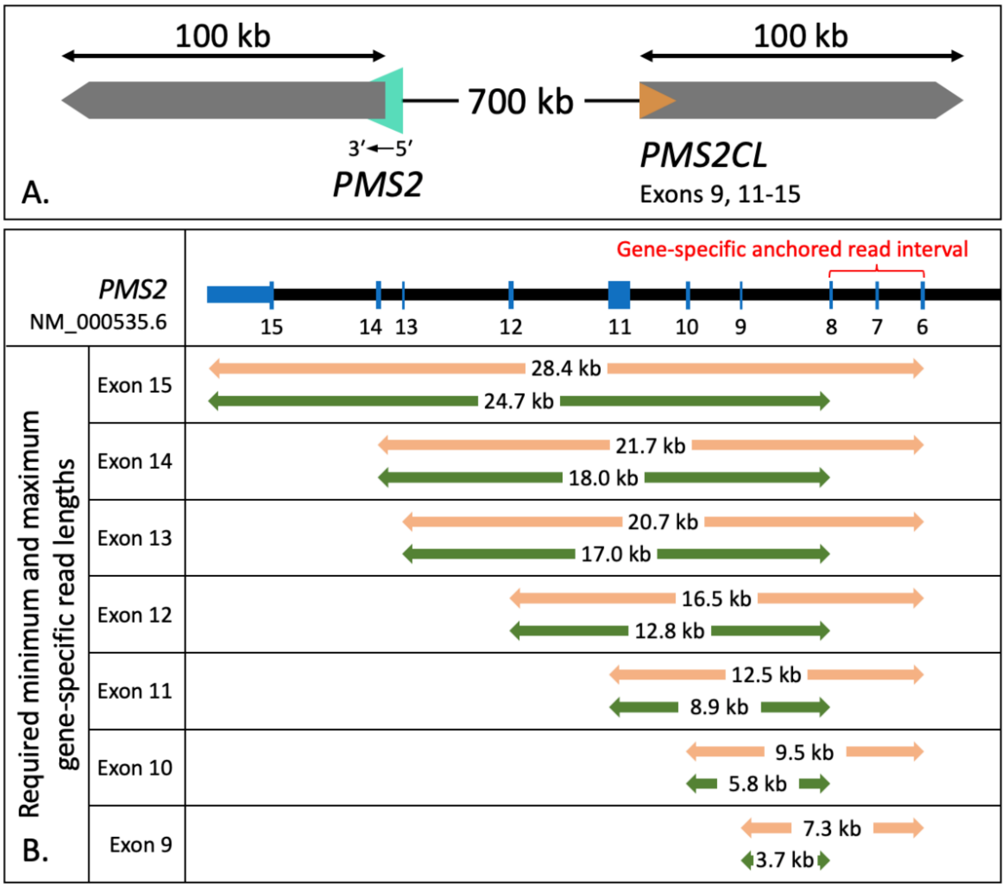

Figure 4.1: A - The genomic structure of the PMS2 duplication. The grey arrows indicate the position of duplicated sequences, while the turquoise and brown arrows show the extent of PMS2 and PMS2CL related sequences respectively. B -The minimum and maximum length of the assembled linked-read molecules required to extend from the gene-specific (non-homologous) region of *PMS2* (exons 6-8) to the 3′ end of the PMS2 gene.

The study aims to use the community curated Genome In a Bottle reference sample NA12878 to first benchmark the performance of the linked-read sequencing in “high-confidence” genomic regions. Having established an appropriate bioinformatics pipeline, the workflow will be used to assay a series of cases that have a confirmed or suspected diagnosis of *PMS2* deficiency.

4.2 Results

4.2.1 Designing a custom hybridisation capture reagent.

An assessment of linked-read technology, by whole genome sequencing, would have been prohibitively expensive at the time the research was undertaken. To develop a cost-effective workflow, it was necessary to ensure enough locus-specific fragments could be sequenced. Whole genome linked-read libraries were therefore enriched using a custom hybridisation capture reagent designed and manufactured by Twist Bioscience. Coordinates for target region of interest were submitted to the Twist design team, who then provided a design report for review prior to manufacture. This comprised 120 nucleotide probes complementary to the target locus; the genomic interval chr7:5391014-7493822 (hg19) which spans a 2,102,808 base pairs region. To avoid capturing off-target regions (sequences originating from elsewhere in the genome), repetitive regions (defined by elements contained in the RepeatMasker track) were excluded. This resulted in 10,577 probes being manufactured which covered 1,269,240 bp (60.36%) of the targeted genomic sequence.

4.2.2 Case selection

To assess the utility of the workflow, for differentiating between pseudogene and non-pseudogene sequences, a diverse group of clinical samples were identified; supporting genotype data was obtained from either the publicly available NA12878 Genome In a Bottle dataset or prior clinical testing carried out by the North-East and Yorkshire Genomic Laboratory Hub. DNA was isolated from either cultured cells, or peripheral blood, using a range of extraction methods. The characteristics of each sample are detailed in Table 4.1.

Table 4.1: Samples selected for analysis by linked-read sequencing.

Sample number	Sex	Reason for selection	DNA extraction method*	Known variant	Exon	Supporting orthogonal dataset
1	M	To evaluate a sample extracted using the laboratories standard protocol.	Chemagic 360	N/A	N/A	None available.
2	F	To use the publicly available NA12878 genotypes to assess the linked-read genotypes in high-confidence regions.	EZ1	N/A	N/A	Publicly available genotypes curated from multiple sequencing technologies.
3	F	To evaluate a structural variant using linked-read technology.	Chemagic 360	*PMS2* whole gene deletion	1-15	Comparative read-depth analysis of hybridisation capture short-read sequencing.
4	M	To assess whether linked-read molecules can confirm whether an exon 15 deletion occurs in the *PMS2* or *PMS2CL* gene.	Unknown	Exon 15 deletion	15	Hybridisation capture enrichment and short-read data in addition to *PMS2* MLPA.
5	M	To assess whether the frameshift variant is in the *PMS2* or *PMS2CL* gene.	Unknown	c.2186_2187del p.(Leu729Glnfs*6	13	Hybridisation capture enrichment and short-read sequencing.
6	M	A positive control for the pathogenic *PMS2* variant.	Unknown	c.2192_2196del p.(Leu731Cysfs*3)	13	Targeted mutation analysis.
7	M	A known pseudogene-derived variant; to determine whether it is located in *PMS2* or *PMS2CL*.	Salt	c.2324A>G p.(Asn775Ser)	14	Hybridisation capture enrichment and short-read sequencing.
8	F	To determine the utility of assessing a *PMS2CL* specific nucleotide which corresponds to the upstream c.2324 position in *PMS2* (to support whether the variant is derived from *PMS2*).	Chemagic 360	c.2350G>A p.(Asp784Asn)	14	Hybridisation capture enrichment and short-read sequencing.

M: Male. F: Female. GIAB: Genome in a bottle (https://www.nist.gov/programs-projects/genome-bottle). Variants reported according to *PMS2* transcript NM_000535.5. N/A: Not applicable. *Sample 2 was extracted from cultured cells; the remaining samples were extracted from peripheral blood.

4.2.3 Sample preparation and GEM formation to create barcoded fragments

To prepare linked-read libraries a low mass of input DNA was required. Genomic DNA samples were first diluted to approximately 20 ng/µl (quantified using the Qubit BR Assay; detection range 0.2-2000 ng/µl). Samples were next diluted to approximately 1 ng/µl and re-quantified (Qubit HS Assay; detection range 5 pg/µl - 120 ng/µl). To ensure the accuracy of the measurement, the samples were assayed 3 times. Diluted samples were determined to be appropriate for linked-read library preparation when the average concentration (n=3) was in the range of 0.8-1.2 ng/µl.

GEM barcoded DNA fragments were created using a Chromium instrument (10X Genomics). To check their size distribution, samples were assessed using a TapeStation (Agilent Technologies) with D5000 ScreenTape (Figure 4.2). Each sample had a size distribution of between 100 to 3,500 base pairs; this was consistent with the manufacturer-specified profile.

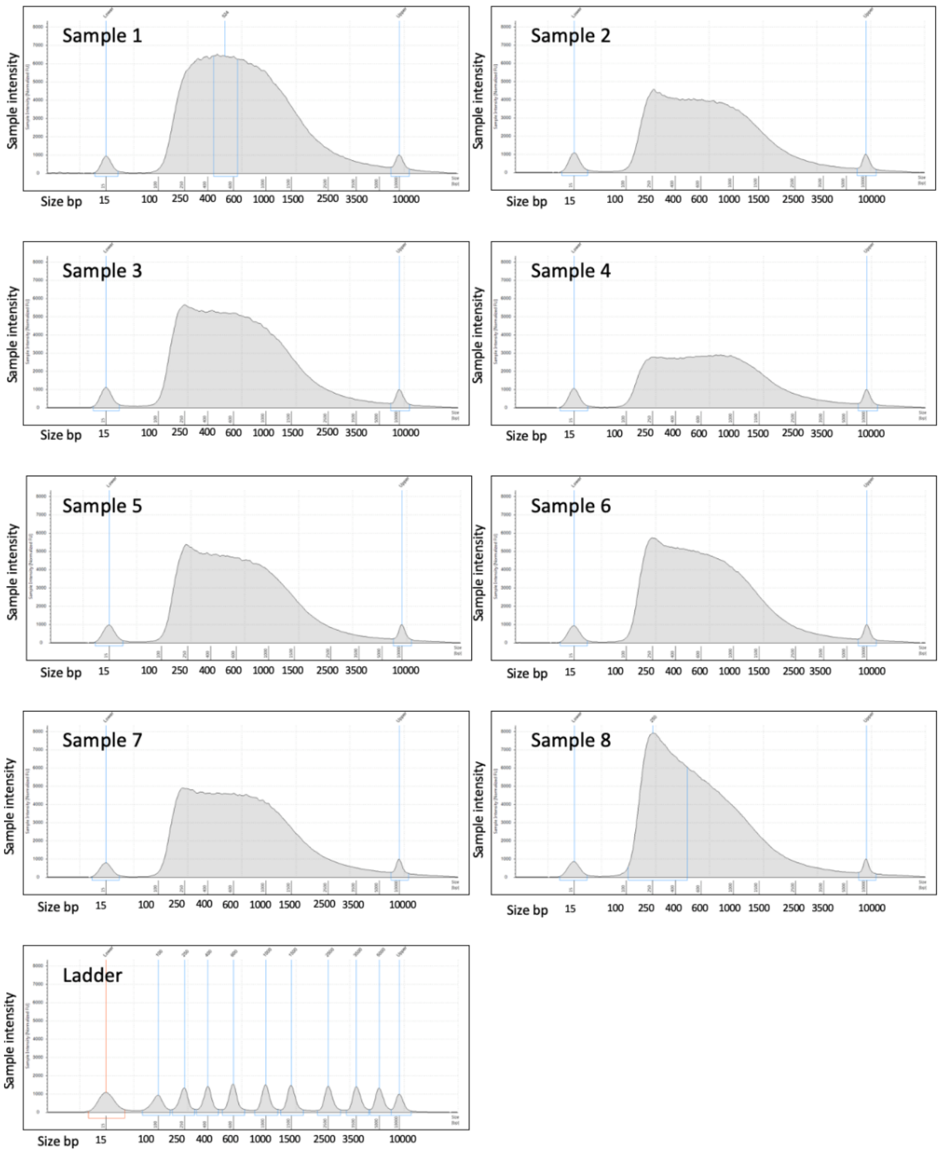

Figure 4.2: The DNA fragment size distribution of each GEM barcoded sample. Sizes are estimated following comparison to fragments of known length determined from the ladder lane. Samples 2, 3, 5, 6, and 7 show comparable profiles. The relative abundance of Sample 4 is lower than the other samples. For each sample the lower and upper marker is highlighted blue and corresponds to fragments of 15-bp and 10,000-bp respectively.

4.2.4 Library preparation, hybridisation capture and sequencing.

GEM barcoded samples were sheared to generate a DNA fragment profile that was suitable for library preparation and subsequent hybridisation capture target enrichment. Sample 1 was processed independently to confirm that the shearing parameters achieved the manufacturer recommended size distribution. Following confirmation of successful shearing, the remaining samples were sheared, with their fragment size distribution profiles shown in Figure 4.3.

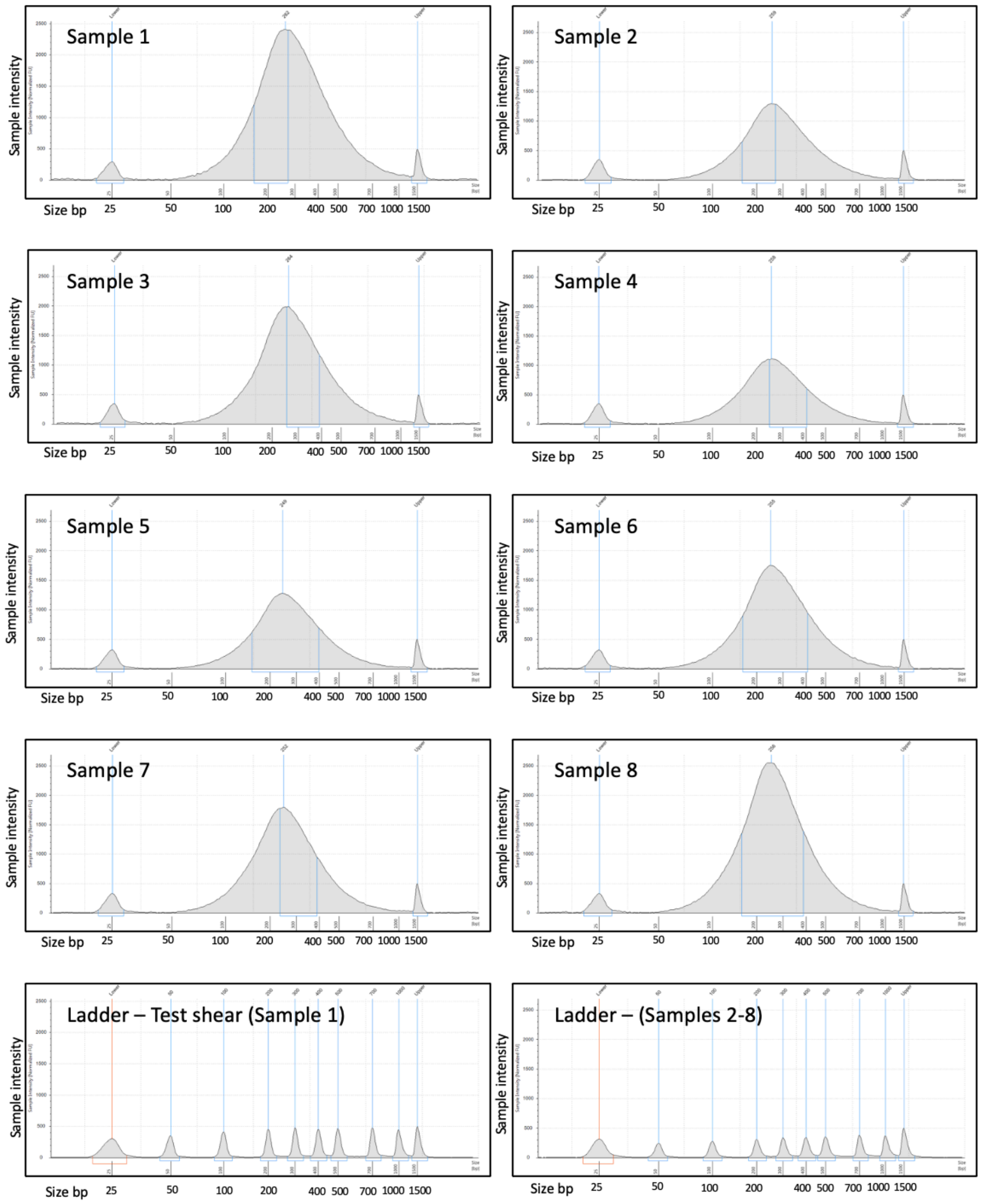

Figure 4.3: The DNA fragment size distribution of sheared GEM barcoded samples. Data were generated using a D1000 ScreenTape. Note the change in the fragment size distribution between the displayed trace and the unsheared profile for the corresponding sample in Figure 4.2. For each sample a lower and upper marker is highlighted blue and corresponds to fragments of 25-bp and 1,500-bp respectively. Sample 1 was processed independently of Samples 2-8 and therefore has a different ladder.

To determine whether adaptor sequences had been ligated to the termini of GEM-barcoded DNA fragments the samples were run on a TapeStation with D1000 ScreenTape (Figure 4.4). The median size distribution increased from approximately 260 bp to approximately 340 bp, confirming that whole genome sequencing libraries had been formed successfully. As the TapeStation data is quantitative this allowed the concentration of each library to be calculated; an equal mass of each library was combined prior to hybridisation capture enrichment (Table 4.2).

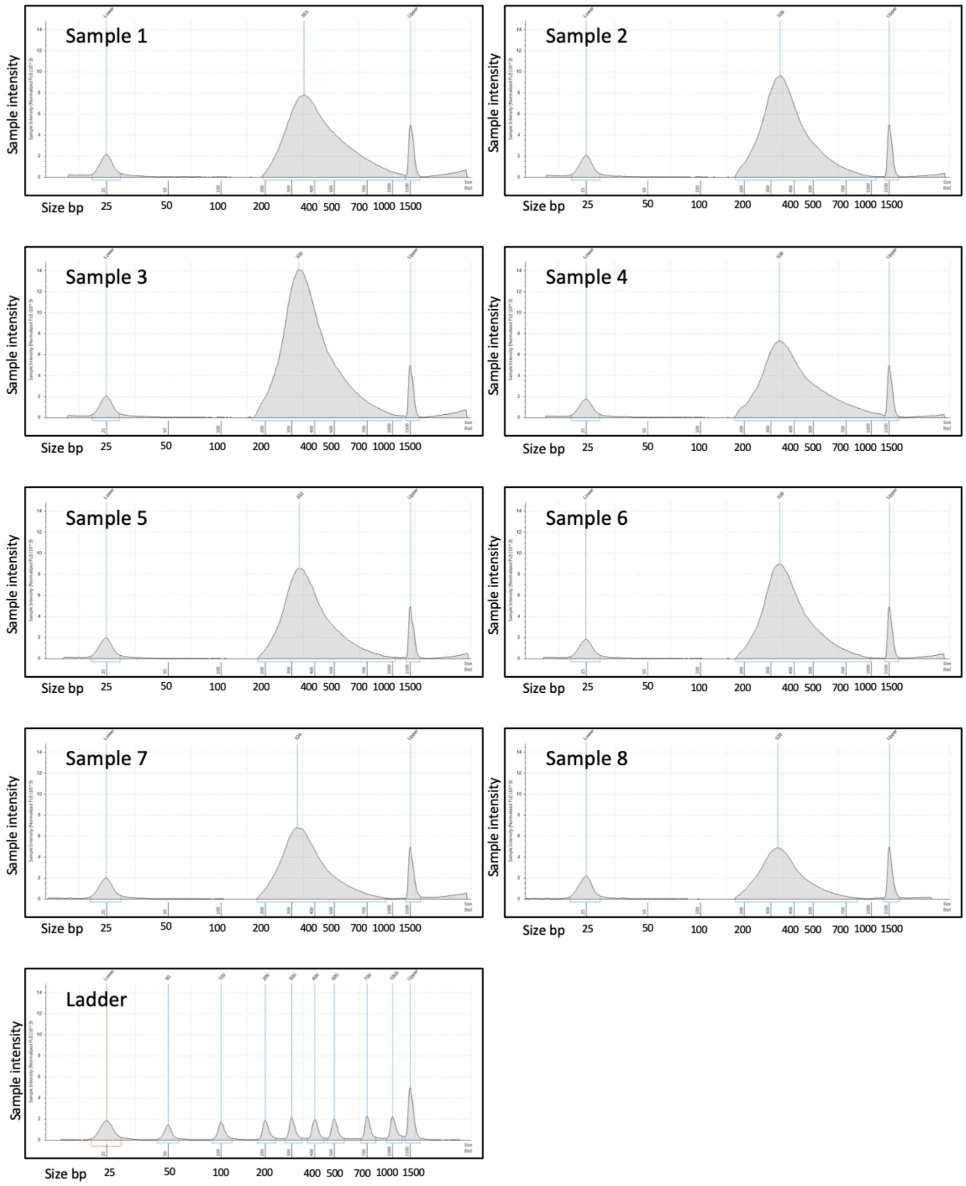

Figure 4.4: The DNA fragment size distribution of GEM barcoded samples following library preparation. Data were generated using a TapeStation with D1000 ScreenTape. Note the increase in fragment size distribution compared to the corresponding pre-adaptor ligation trace in Figure 4.3. For each sample a lower and upper marker is highlighted blue and corresponds to fragments of 25-bp and 1,500-bp respectively.

Table 4.2: The concentration and volume of library required for equimolar pooling.

Sample number	Library concentration* (ng/µl)	Volume for 375 ng (µl)
1	102.0	3.68
2	111.0	3.38
3	169.0	2.22
4	96.8	3.87
5	96.3	3.89
6	107.0	3.50
7	72.0	5.21
8	53.1	7.06

*Assayed using a TapeStation with D1000 ScreenTape.

Whole genome libraries were enriched at the target locus. To verify that the fragment size distribution had been maintained, the post-capture pool was assessed using a TapeStation with D1000 high-sensitivity ScreenTape. The final pool of sequencing libraries had a peak size of approximately 340 bp (Figure 4.5) and concentration of 8.51 ng/µl (assayed by Qubit); the molarity of the pool, based on average size and Qubit quantification, was 39 nM.

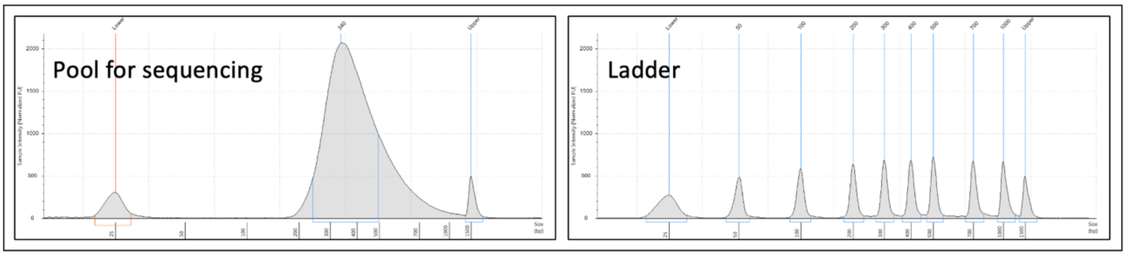

Figure 4.5: The DNA fragment size distribution of the combined GEM-barcoded libraries following hybridisation capture enrichment. The limited mass necessitated use of a high-sensitivity D1000 ScreenTape.

The libraries were sequenced on a NextSeq 500 using a High Output cartridge configured to generate paired-end 151 bp reads. The denatured pool was loaded at a molarity of 1.6 pM. The average cluster density of the flow cell was 190 k/mm^2 with 92.7% of clusters passing the sequencer filtering criteria. The run yielded 125 gigabases of raw sequence data with 83.9% of bases having a Phred-scaled quality score ≥Q30. These metrics are within the manufacturer specified parameters (100-120 Gb; >75% of bases ≥Q30) and were indicative of a successful sequencing run. Raw data were demultiplexed by sample index tag and converted to FASTQ.gz format. The resulting per-sample read distribution is reported in Table 4.3. For a pool of 8 samples the expected proportion of reads per library, following perfect equimolar pooling, was 12.5%. While the distribution of reads between samples is comparable to that seen in non-linked read datasets, it is notable that a lower number of reads was observed for Sample 4.

Table 4.3: The total number of sequenced inserts and read distribution per library.

Sample number	Sequenced inserts*	Proportion of total reads (%)
1	48,686,976	10.62
2	54,180,127	11.82
3	58,236,355	12.71
4	38,934,237	8.49
5	66,770,466	14.57
6	64,767,701	14.13
7	59,426,084	12.97
8	67,354,742	14.69
Total	458,356,688	100.00

*Also termed “read pairs”, these values correspond to the number of pairs of forward and reverse sequence reads.

4.2.5 Establishing a Longranger linked-read alignment pipeline.

To demultiplex the pool of sequenced libraries, and construct linked-read information, the proprietary Longranger Targeted Pipeline was deployed (Figure 2.1). The data processing workflow enabled the barcoded GEM-bead sequence reads to be aligned to the human reference genome (build hg19). This meant sequence reads originating from the same source DNA fragment could be identified and marked with the unique molecular barcode sequence.

Alignment performance metrics are reported in Table 4.4. For each library, a high proportion (≥94%) of sequence reads were mapped to the reference genome which is comparable to non-linked read datasets generated in our laboratory. Nevertheless, the proportion of sequence “on-target” reads that aligned to the genomic region that was targeted by the custom hybridisation probes was low (mean: 14.72%), suggesting suboptimal probe hybridisation and/or inefficient removal of non-specific binding at the capture stage, leading to amplification of greater off-target product at the post-capture enrichment stage. The proportion of on-target reads that would be expected from a typical hybridisation capture experiment that used Twist Biosciences capture probes in combination with the manufacturer's whole genome library preparation reagent is expected to be approximately 90% (Marosy, 2018). Consistent with Sample 4 producing the lowest total number of sequencing reads, this sample also had the lowest proportion of on-target reads (13.86%).

The duplicate rate, the proportion of sequence reads that have arisen due to PCR amplification rather than from an independent DNA molecule, was on average 24.2%; this is higher than would be expected from libraries captured using a non-linked read whole genome library preparation. Causes of high rates of duplicate reads in NGS can include low quantities of input DNA for the initial library preparation, leading to PCR duplicates created during pre-capture enrichment, or low yield from the capture of the hybridised targets, leading to duplicates from the final enrichment PCR. Reported duplicate rates for non-linked read libraries, both in the literature and from our molecular diagnostic workflows, are typically 2-8% (Marosy, 2018). In addition to the Sample 4 library having the lowest proportion of on target reads, it also had a slightly higher duplicate rate (26.9%). The low input mass of DNA required to generate GEM-barcoded whole genome libraries, in combination with the PCR cycles required to achieve the high mass input required for hybridisation capture enrichment, likely explains the source of the high duplicate rate.

Table 4.4: Alignment performance metrics for each sample after processing using the Longranger pipeline.

Sample number	Reads mapped to the reference genome (%)	Reads mapped to target region* (%)	Proportion of reads classed as duplicates (%)
1	97.1	15.30	23.6
2	96.5	15.08	23.5
3	97.5	14.90	20.7
4	94.0	13.86	26.9
5	96.2	14.56	24.0
6	96.9	15.18	22.9
7	95.2	14.51	24.5
8	96.1	14.36	27.3
Mean	96.2	14.72	24.2

*Defined by the genomic coordinates chr7:5391014-7493822 (Human reference genome build hg19).

The relationship between individual sequence reads and the DNA fragment from which they originated is defined by a “linked-read” barcode. This barcode is stored as a “BX tag” in the aligned bam file and can be used in downstream data processing to ascertain the molecular origin of the reads. The performance of each linked-read library can be assessed from the metrics detailed in Table 4.5. Each GEM contains a uniquely barcoded bead droplet, and the unique barcode is incorporated into the DNA fragment contained within the droplet. While it is important to have a sufficient number of recovered GEMs to sequence the locus, it is the molecule length that is of particular relevance to enabling unambiguous alignment to either *PMS2* or its pseudogene *PMS2CL*. Consistent with previous observations, the library for Sample 4 was an outlier for the linked-read metrics. For this library, while a large number of GEMs were detected, the amount of DNA sequenced per GEM was low. This likely accounts for the comparatively short total molecule length of reconstructed fragments.

As a bioinformatic quality assurance step, the number of reported total reads (forward and reverse sequences) determined by Longranger was compared to the number of sequenced inserts. As expected, the Longranger identified total read count was twice the number of sequenced inserts.

Table 4.5: Linked-read performance metrics for each sample after processing using the Longranger pipeline.

Sample number	GEMs detected	Mean DNA per GEM (bp)	Mean length of reconstructed molecules (bp)	Total reads
1	34,357	33,799,589	33,553	97,373,882
2	59,610	16,904,030	19,196	108,360,180
3	36,477	29,950,669	37,350	116,472,650
4	156,829	5,302,536	10,730	77,868,414
5	79,534	16,688,037	17,947	133,540,830
6	46,382	26,619,624	28,863	129,535,334
7	86,644	10,974,138	15,082	118,852,074
8	52,648	13,384,727	26,231	134,709,350
Mean	69,060	19,202,919	23,619	114,589,089

4.2.6 Assessment of hybridisation capture linked-read library preparation as an alternative enrichment approach.

Sample 2 is the GIAB reference sample (NA12878); this is a publicly available specimen that has been comprehensively genotyped using a range of orthogonal technologies. To assess the performance of hybridisation capture linked-read enrichment, as a generic library preparation and enrichment workflow, the genotypes identified in Sample 2 were compared to the freely available reference dataset for this sample.

In short read sequencing, detection of structural variation can be challenging due to reduced mapping context when aligning data to the reference genome. The higher sensitivity of mutation detection for single nucleotide variants (SNVs) therefore prompted this class of variant to be analysed separately from the insertion-deletion variants (indels). The GIAB variant call format (VCF) file (which contains machine readable variant information) was filtered to include only “high confident” intervals spanning unique regions of the target locus; this extended between chr7:6049657-6763395 (hg19) and included 332 genomic region (649 kb of sequence).

An automated pairwise comparison between the reference genotypes and those identified from the linked-read data was performed using the bioinformatic tool vcfeval. Concordant (matching) genotypes were identified for 587 SNVs and 86 indels resulting in sensitivity values of 97.83% and 80.37% respectively (Table 4.6). As anticipated, the number of false positive and false negative variant calls was greatest for indel variants. To investigate these discrepancies, raw sequence reads were visualised using the IGV and the basis for each erroneous (non-matching) variant call was determined.

Table 4.6: Performance of the Longranger variant calling workflow for the GIAB NA12878 sample.

Variant type	TP call	FP call	FN call	Sensitivity (%)
Single nucleotide variant	587	4	13	97.83
Insertion-deletion variant	86	46	21	80.37

TP: True positive, FP: False positive, FN: False negative.

Manual scrutiny of the raw sequence reads revealed that relatively few scenarios accounted for the discrepant SNV genotypes (Table 4.7). There were two false positive variant calls (at positions chr7: 6290974 and chr7:6403076) that would have been discounted following end-user review, and a further two false positive discordant genotypes where the incorrect assignment of zygosity could easily be resolved.

Of the 13 instances of false negative (missed) SNV calls 6 were due to the variants being flagged, and subsequently removed, due to a “phasing filter” included in the pipeline; visual inspection of the data confirmed the variants were detectable by the assay. Relaxing the parameters of this filter may have allowed the variants to be identified in a future iteration of the pipeline. A further two genotypes were discordant due to the incorrect assignment of zygosity (a homozygous rather than a heterozygous, and a heterozygous rather than a homozygous variant call). This inappropriate assignment was readily visible following manual inspection of the data and adjustments to the pipeline may enable improvement to variant calling at these loci, but possibly at the expense of incorrect calling elsewhere. The final five variants were refractory to detection due to there being insufficient sequence reads to a make an accurate assessment of the queried genomic region, this is a limitation of the assay at these loci.

Table 4.7: Manual scrutiny of discordant single nucleotide variant calls in the 10X data for GIAB NA12878 sample compared to the reference data set.

Type	Class	Position	Ref	Alt	T-GT	Explanation
SNV	FP	6290974	C	G	N/A	Variant called at end of 14 nt poly(T)
SNV	FP	6403076	A	T	N/A	Low VAF (T=0.15) in a TTTAA repeat
SNV	FP	6552066	A	G	Het	Incorrect zygosity (called as Hom) VAF: A=0.23 G=0.77
SNV	FP	6716546	C	A	Hom	Incorrect zygosity (called as Het) VAF: C=0.03 A=0.97
SNV	FN	6111534	A	G	Het	Failed phasing filter, sufficient depth
SNV	FN	6327158	C	A	Het	Insufficient coverage (2x total, 2 Alt)
SNV	FN	6327221	G	A	Het	Insufficient coverage (1x total, 1 Alt)
SNV	FN	6359505	A	T	Het	Insufficient coverage (13x total, 4 Alt)
SNV	FN	6419049	T	G	Het	Insufficient coverage (4x total, 1 Alt)
SNV	FN	6504240	T	C	Het	Insufficient coverage (8x total, 0 Alt)
SNV	FN	6552066	A	G	Het	Incorrect zygosity (called as Hom) VAF: A=0.23 G=0.77
SNV	FN	6600373	G	T	Het	Failed phasing filter, sufficient depth
SNV	FN	6611882	G	A	Het	Failed phasing filter, sufficient depth
SNV	FN	6686601	T	C	Het	Failed phasing filter, sufficient depth
SNV	FN	6716546	C	A	Hom	Incorrect zygosity (called as Het) VAF: C=0.03 A=0.97
SNV	FN	6717013	T	C	Het	Failed phasing filter, sufficient depth
SNV	FN	6762985	A	G	Het	Failed phasing filter, sufficient depth

Chromosome 7 position is provided according to human reference genome hg19. T-GT: True genotype. N/A: Not applicable. Het: Heterozygous. Hom: Homozygous. FP: False positive. FN: False negative. VAF: Variant allele frequency. nt: Nucleotide. del: Deletion. ins: Insertion.

By contrast to SNVs, a greater number of indel variant false positive (46 instances in total) and false negative (21 instances in total) calls were identified (Table 4.8). Manual scrutiny of the sequence reads revealed that the majority of false positive calls (43 instances) were identified at the boundaries of poly(N) tracts and included the erroneous insertion or deletion of 1 or more nucleotides. The lengths of these tracts ranged from 7 to 22 nucleotides. A further two false positive calls occurred at dinucleotide (AT or AC) repeat sequences. The limitation of being unable to correctly genotype sequences at, and adjacent to, poly(N) tracts by next-generation sequencing has been a widely reported. The final false positive variant call (at position chr7:6495175) intersected a low-complexity AluY SINE element. The variant allele frequency (0.25) and complex genomic architecture of this region suggested this to be an erroneous call.

False negative indel variant calls (21 instances in total) were characterised by their proximity to poly(N) tracts. At 5 positions the length or zygosity of a poly(N) tract was incorrectly determined. A further 7 variants were flagged as having failed a bioinformatic quality control filter (either related to phasing or homopolymers). For 4 variants, while “soft-clipped” reads supported the presence of the variant, the variant calling pipeline did not identify these locations. While adjustments to the pipeline parameters may improve the sensitivity to detect these variants, their proximity to poly(N) tracts will likely mean that their analysis remains difficult. A final group of 5 indel variants had insufficient read coverage to permit their analysis.

Table 4.8: Manual scrutiny of discordant insertion-deletion variant calls in the 10X data for GIAB NA12878 sample compared to the reference data set.

Type	Class	Position	Ref	Alt	T-GT	Explanation
INDEL	FP	6089258	CT	C	Het	Hom del at 17 nt poly(T)
INDEL	FP	6129167	CT	C	N/A	Het del at 14 nt poly(T)
INDEL	FP	6146362	AT	A	N/A	Het del at 14 nt poly(T)
INDEL	FP	6181852	C	CT	N/A	Het ins at 13 nt poly(T)
INDEL	FP	6192933	CA	C	N/A	Het del at 12 nt poly(A)
INDEL	FP	6208594	T	TC	N/A	Het ins at 7 nt poly(C)
INDEL	FP	6247580	T	TG	N/A	Het ins at 7 nt poly(G)
INDEL	FP	6275831	A	AC	N/A	Het ins at 8 nt poly(C)
INDEL	FP	6292847	A	AT	N/A	Het ins at 18 nt poly(T)
INDEL	FP	6314338	T	TG	N/A	Het ins at 7 nt poly(G)
INDEL	FP	6375803	C	CA	Het CAA	Het ins CA rather than Het ins CAA at 16 nt poly(A)
INDEL	FP	6376313	CA	C	N/A	Het del at 16 nt poly(A)
INDEL	FP	6383028	CATAT	C	Het CAT	Hom del CATAT rather than Het del CAT at AT repeat sequence
INDEL	FP	6402408	TA	T	N/A	Het del at 12 nt poly(A)
INDEL	FP	6430238	AT	A	N/A	Het del at 10 nt poly(A) rather than het SNV at adjacent nt
INDEL	FP	6430851	C	CT	N/A	Het ins at 13 nt poly(T)
INDEL	FP	6468117	CT	C	N/A	Het del at 12 nt poly(T)
INDEL	FP	6470058	C	CT	N/A	Hom ins at 10 nt poly(T)
INDEL	FP	6489238	TTA	T	N/A	TA del at 9 nt poly(A) rather than het SNV at adjacent nt
INDEL	FP	6494120	CAA	C	Het CA	Het del at 22 nt poly(A)
INDEL	FP	6495175	GGGAC	G	N/A	Spurious false positive (VAF=0.25)
INDEL	FP	6495453	C	CA	N/A	Het ins at 9 nt poly(A)
INDEL	FP	6594654	CA	C	N/A	Het del at 15 nt poly(A)
INDEL	FP	6601642	GA	G	N/A	Het del at 15 nt poly(A)
INDEL	FP	6605654	C	CT	N/A	Het ins at 11 nt poly(T)
INDEL	FP	6608086	CA	C	N/A	Het del at 14 nt poly(A)
INDEL	FP	6611064	CA	C	N/A	Het del at 14 nt poly(A)
INDEL	FP	6612706	AT	A	N/A	Het del at 16 nt poly(T)
INDEL	FP	6614202	GA	G	N/A	Het del at 15 nt poly (A)
INDEL	FP	6618836	CT	C	N/A	Het del at 13 nt poly(T)
INDEL	FP	6622808	CA	C	N/A	Het del at 13 nt poly(A)
INDEL	FP	6623679	GA	G	N/A	Het del at 13 nt poly(A)
INDEL	FP	6630475	C	CT	N/A	Het ins at 11 nt poly(T)
INDEL	FP	6656368	GA	G	N/A	Het del at 14 nt poly(A)
INDEL	FP	6683904	T	TG	N/A	Het ins at 7 nt poly(G)
INDEL	FP	6688690	T	TG	N/A	Het ins at 8 nt poly(G)
INDEL	FP	6697914	C	CT	N/A	Het ins at 10 nt poly(T)
INDEL	FP	6720868	C	CT	N/A	Het ins at 10 nt poly(T)
INDEL	FP	6723575	CT	C	N/A	Het del at 13 nt poly(T)
INDEL	FP	6726731	C	CA	N/A	Het ins at 13 nt poly(A)
INDEL	FP	6729528	G	GA	N/A	Het ins at 9 nt poly(A)

INDEL	FP	6732517	AAC	A	N/A	Het AC del within AC repeat
INDEL	FP	6736600	T	TA	N/A	Het ins at 12 nt poly(A)
INDEL	FP	6741582	CA	C	N/A	Het del at 20 nt poly(A)
INDEL	FP	6743198	T	TG	N/A	Het ins at 7 nt poly(G)
INDEL	FP	6743946	TA	T	N/A	Het del at 12 nt poly(A)
INDEL	FN	6089258	CT	C	Het	Hom del rather than het del at 17 nt poly(T)
INDEL	FN	6108392	CA	C	Hom	Failed phasing filter, at 17 nt poly(A)
INDEL	FN	6131311	T	TACACACACACAC	Het	Failed phasing filter, at AC repeat tract
INDEL	FN	6309219	A	AGTAGAGAC	Het	Soft-clipped reads supporting variant visible; no variant call
INDEL	FN	6333218	CT	C	Het	Failed phasing filter, at 15 nt poly(T)
INDEL	FN	6337963	T	TTTTA	Het	Insufficient coverage (7x total, 0 Alt)
INDEL	FN	6375803	C	CAA	Het	Het ins CA rather than CAA, at 16 nt poly(A)
INDEL	FN	6383028	CATATAT	CAT	Het	Hom del rather than Het del, at AT repeat tract
INDEL	FN	6408869	C	CA	Het	Insufficient coverage (12x total, 0 Alt)
INDEL	FN	6412423	CTTT	C	Het	Soft-clipped reads supporting variant visible; no variant call
INDEL	FN	6421021	CTT	C	Het	Failed phasing filter, at 17 nt poly(T)
INDEL	FN	6493180	C	CAA	Het	Soft-clipped reads supporting variant visible; no variant call
INDEL	FN	6494120	CAA	CA	Het	Het del AA rather than Het ins A, at 22 nt poly(A)
INDEL	FN	6504365	CT	C	Hom	Insufficient coverage (9x total, 0 Alt)
INDEL	FN	6562438	A	ATGTGTGTGTGTGTG	Het	Insufficient coverage (8x total, 1 Alt)
INDEL	FN	6586200	C	CA	Hom	Het ins A rather than Hom ins A, at 17 nt poly(A)
INDEL	FN	6595299	CT	C	Het	Insufficient coverage (10X total, 5 Alt)
INDEL	FN	6598458	C	CA	Hom	Failed phasing filter, at 17 nt poly(A)
INDEL	FN	6610761	T	TC	Het	Failed homopolymer filter
INDEL	FN	6610768	C	CA	Hom	Soft-clipped reads supporting variant visible; no variant call
INDEL	FN	6740583	ATTTGTTCG	A	Het	Failed phasing filter

Chromosome 7 position is provided according to human reference genome hg19. T-GT: True genotype. N/A: Not applicable. Het: Heterozygous. Hom: Homozygous. FP: False positive. FN: False negative. VAF: Variant allele frequency. nt: Nucleotide. del: Deletion. ins: Insertion.

4.2.7 Establishing a "truth-set" of *PMS2* genotypes from long-range PCR amplification products.

Long-range PCR amplification products were generated for Samples 1 and 2 using a primer that bound to a unique region of *PMS2*. The longer, so-called "Edinburgh" assay, amplified from intron 10 to exons 15 while the shorter, so called "Set 3" assay amplified from exon 10 to intron 12. The cumulative coverage profiles and genotypes identified using these *PMS2*-specific amplicons are displayed in Figure 4.6. For both samples the variants identified in the regions of overlap were concordant. For Sample 2 (the GIAB reference sample), genotypes were compared to the publicly available dataset (Table 4.9). Of the 19 variants identified, 9 were concordant with the public dataset. The remaining 10 variants were heterozygous and discordant, though it was noted that these were flagged as "low confidence" genotypes in the public data. All genotypes that were identified in the long-range PCR dataset were identified and concordant with the linked-read data for these samples.

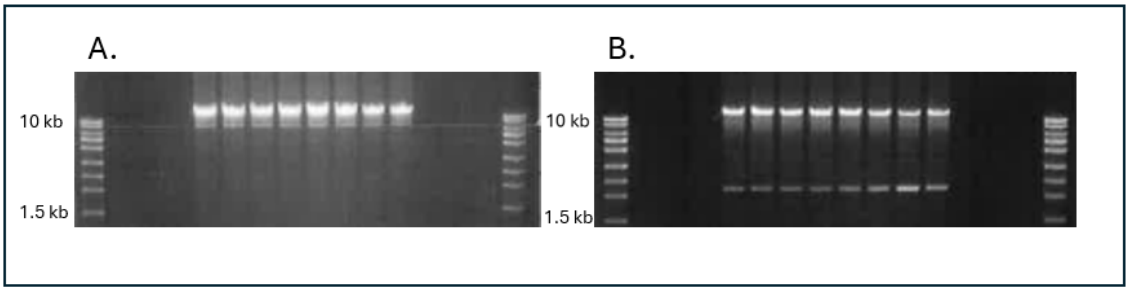

Figure 4.6: PCR products generated from control samples using the "Edinburgh" assay primers for PMS2 exons 11-15. Ladder used is High Ranger from Norgen Biotech. **(A)** Sample 1 normal male control. **(B)** Sample 2 NA12878 GIAB sample.

Table 4.9: Variants identified in long-range PCR amplicons for Samples 1 and 2.

Sample	Variant ID*	g.Nomen	Zygosity	Concordant between amplicons	Concordant with public genotype	Identified in linked-read data
1	1	6029012T>C	Het	Y	N/A	Y
	2	6028768G>A	Het	Y		Y
	3	6027702A>C	Het	Y		Y
	4	6026775T>C	Het	Y		Y
	5	6025894C>T	Het	Y		Y
	6	6024035G>A	Het	Y		Y
	7	6019224T>A	Hom	N/A		Y
	8	6017902G>A	Het			Y
	9	6016470C>A	Het			Y
	10	6016109A>G	Het			Y
2 (NA12878)	1	6029012T>C	Hom	Y	Y	Y
	2	6028768G>A	Hom	Y	Y	Y
	3	6026775T>C	Hom	Y	Y	Y
	4	6025980G>A	Het	Y	N (LC region)	Y
	5	6025894C>T	Het	Y	Y	Y
	6	6024053G>A	Het	Y	N (LC region)	Y
	7	6024035G>A	Het	Y	N (LC region)	Y
	8	6023033T>C	Het	Y	Y	Y
	9	6022961C>T	Het	Y	Y	Y
	10	6022866C>T	Het	Y	Y	Y
	11	6019224T>A	Hom	N/A	Y	Y
	12	6017523A>G	Het		N (LC region)	Y
	13	6016470C>A	Het		N (LC region)	Y
	14	6016109A>G	Hom		Y	Y
	15	6014998A>T	Het		N (LC region)	Y
	16	6014988T>C	Het		N (LC region)	Y
	17	6014512T>C	Het		N (LC region)	Y
	18	6013851_6013852delinsAG	Het		N (LC region)	Y
	19	6013049C>G	Het		N (LC region)	Y

*See Figure 4.6 for corresponding IGV plots. Coordinates provided for chromosome 7 according to build hg19 of the human reference genome. Het: Heterozygous. Hom: Homozygous. N/A: Not applicable. LC: Low confidence.

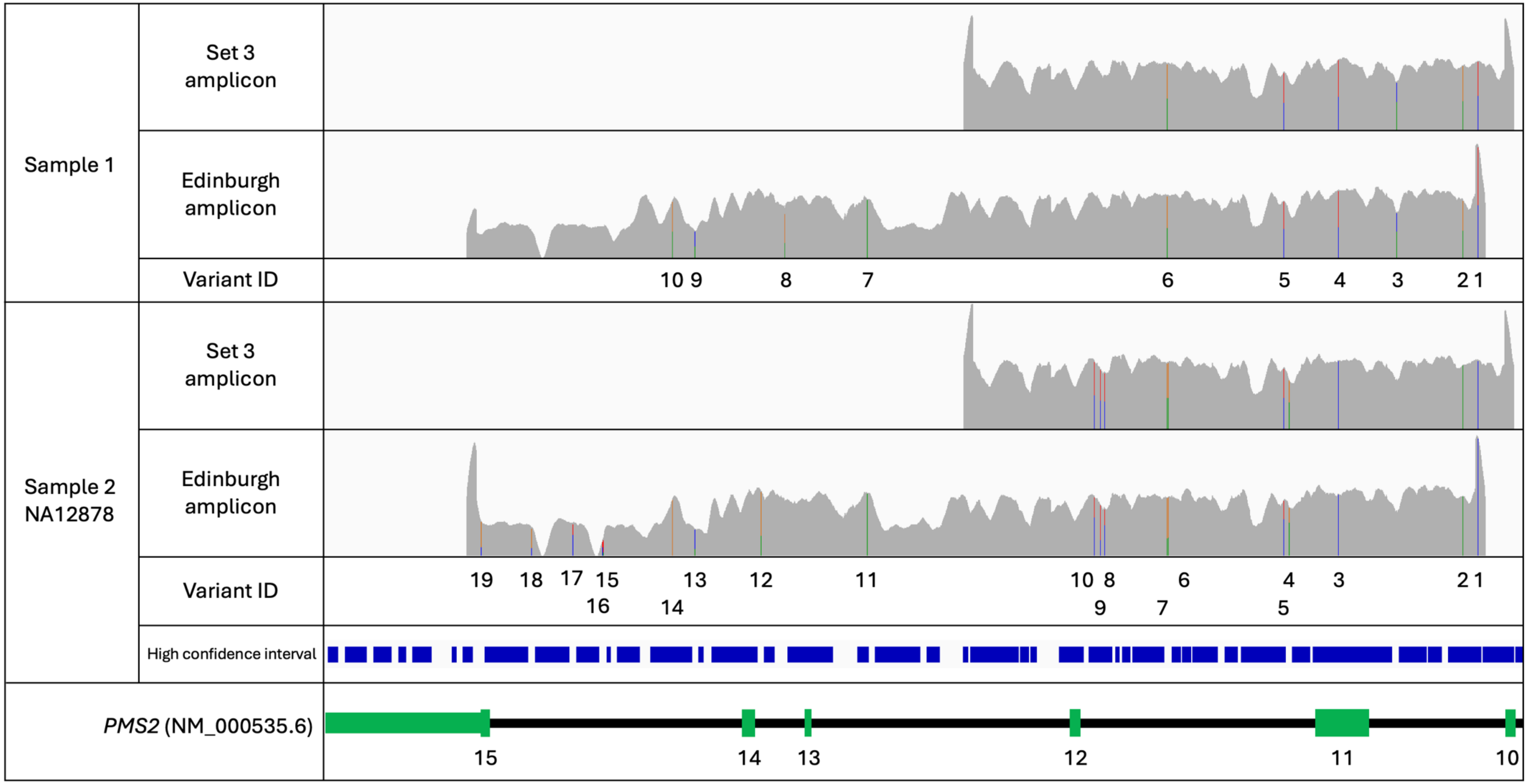

Figure 4.7: Genotypes identified following long-range PCR amplification of Samples 1 and 2. Owing to the high sequencing yield, datasets were down sampled to 0.05 of total reads. The *y*-axis of the cumulative coverage plot was 0-13,000× and 0-7,000× for the Set 3 and Edinburgh assays respectively. The region defined as "high confidence" in the public dataset for NA12878 genotypes is displayed in the high confidence interval track.

4.2.8 Detection and interpretation of *PMS2/PMS2CL* variants in clinical specimens

To assess the utility of linked-read sequencing to characterise variants in either *PMS2* or *PMS2CL,* a series of cases with known or suspected *PMS2*/*PMS2CL* genotypes were sequenced.

4.2.8.1 Sample 3

The female patient was referred for testing following a diagnosis of bowel cancer at aged 57. There was a notable family history with her father having bowel cancer at aged 55. Analysis of microsatellite markers generated a microsatellite unstable result, consistent with a diagnosis of Lynch syndrome. Standard-of-care genetic testing of the coding and immediate splice site sequences of Lynch syndrome associated genes (MLH1, MSH2, MSH6 and PMS2) was performed by hybridisation capture and short-read sequencing. Comparative read depth analysis of these data revealed a heterozygous multi-exon *PMS2* deletion; while the proximal boundary of the deletion was predicted to extend upstream of *PMS2* exon 1, assessment of the distal deletion breakpoint was impaired by ambiguous mapping across the homologous *PMS2* exons.

CNVseq (low-coverage short-read whole genome sequencing) was performed to further characterise the deletion. This yielded 21,455,951 single end 76-bp reads which were processed using the in-house pipeline described in 2.8.2. The heterozygous deletion was estimated to be 88 kb in size, with predicted deletion breakpoints extending between chr7:5980954-6069411; this encompassed the entire *PMS2* gene (Figure 4.8). Nevertheless, low depth WGS was insufficient to enable the deletion breakpoint to be characterised at single nucleotide resolution, prompting its analysis by linked-read hybridisation capture enrichment.

The probes included in the hybridisation capture reagent extended beyond the putative deletion site. Linked reads were aligned and subsequently separated into their phased haplotypes. This defined the distal deletion breakpoint within *CCZ1* intron 1 (Figure 4.9A). The absence of reads at the proximal boundary suggested the breakpoint occurred within *EIF2AK1* intron 8. However, further scrutiny of the sequence reads prompted by a "square cut" coverage profile revealed soft-clipped breakpoint spanning sequence reads in *EIF2AK1* intron 7. The breakpoint spanning reads mapped to the opposite strand, suggesting the sequence had been inverted. The deletion and inverted breakpoint were confirmed by Sanger sequencing of a PCR amplicon that spanned the mutant allele (Figure 4.9B). The intersected sequence *EIF2AK1* sequence spanned a low-complexity repeat defined as a MER58B element in the RepeatMasker track on the UCSC Genome Browser (Perez *et al.,* 2025). A literature search of the additional genes that were deleted at this locus did not reveal any further known disease-associated phenotypes.

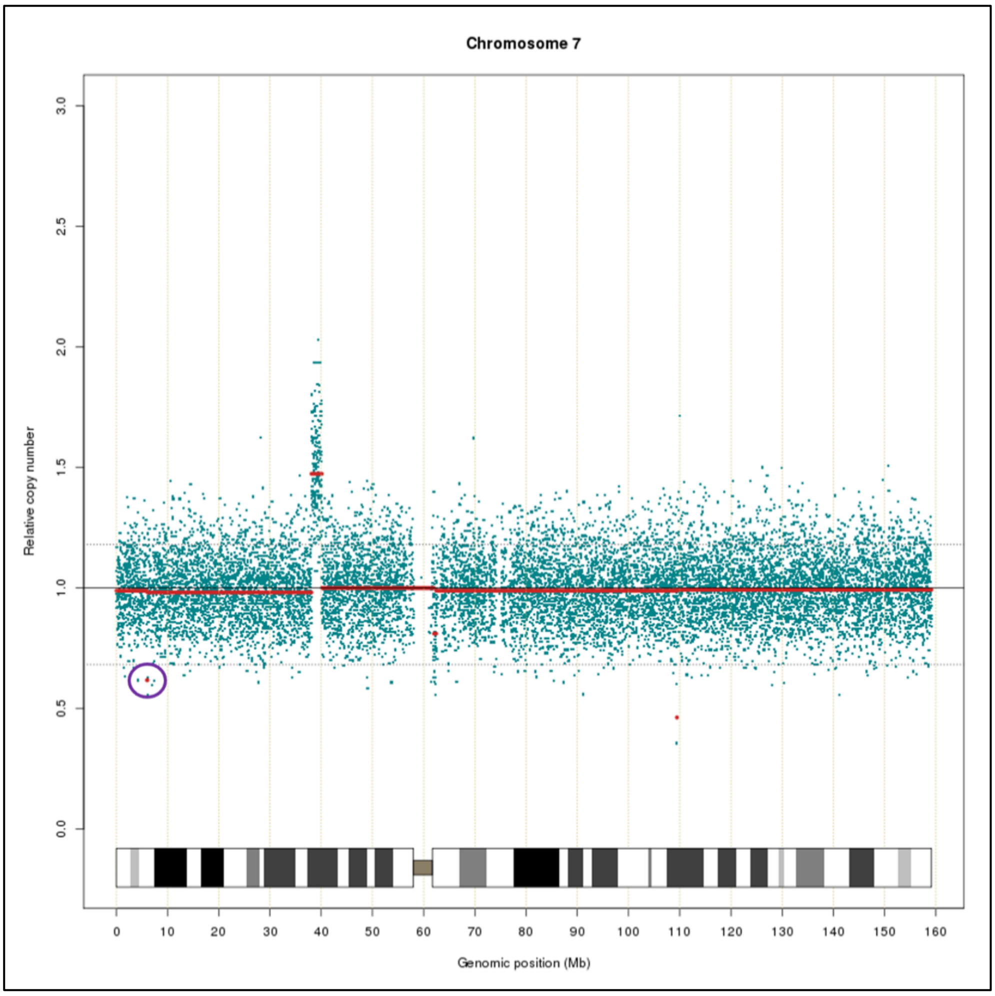

Figure 4.8: Chromosome CNVseq karyogram for Sample 3. A single copy (heterozygous) deletion encompassing *PMS2* is marked by the purple circle (chr7:5980954-6069411). A 2 Mb single copy gain (chr7:38108790-40122506) and 15.5 kb single copy deletion (chr7:109435829-109452361) did not contain disease-associated genes that were consistent with the clinical phenotype.

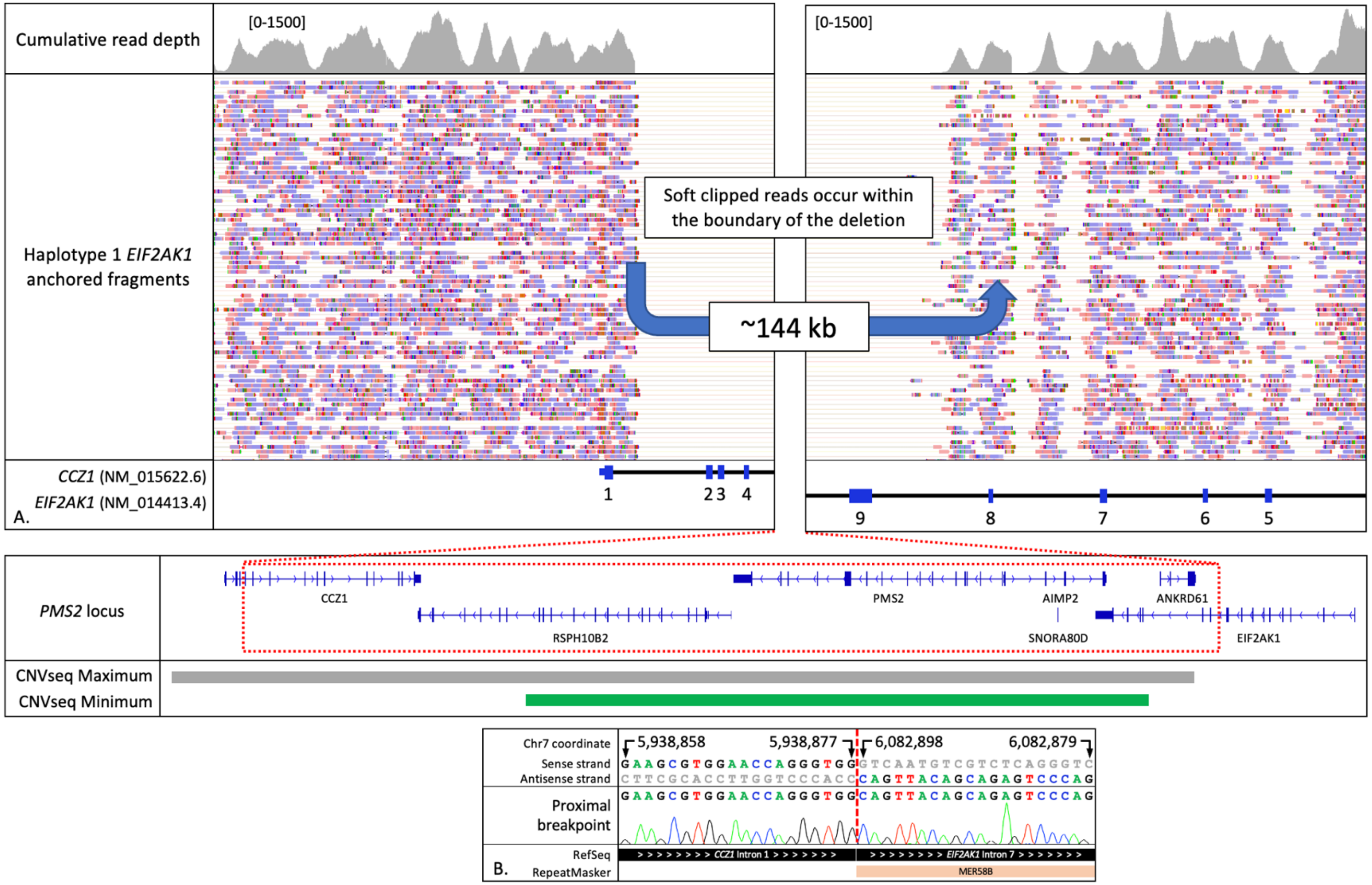

Figure 4.9: Single nucleotide characterisation of a heterozygous PMS2 whole-gene deletion in sample 3. (A) Linked read alignments were separated into their constituent phased haplotypes. This revealed a region of deleted sequence and a cluster of soft-clipped reads within *EIF2AK1* intron 7. Manual scrutiny of the soft-clipped reads suggested the sequence was inverted. **(B)** The precise breakpoint was characterised by breakpoint spanning PCR and Sanger sequencing.

4.2.8.2 Sample 4

A male patient was diagnosed with colon cancer at age 40. His family history included his father being affected with colorectal cancer at age 61 and his brother being affected with a tubular adenoma with low grade dysplasia at the age of 62. Analysis of the patient's tumour revealed immunohistochemical loss of PMS2 staining and microsatellite instability. The patient underwent germline testing by hybridisation capture and short-read sequencing which revealed an apparent deletion of *PMS2* exon 15. Multiplex ligation-dependent probe amplification was performed by the diagnostic genetics lab for sample 4 using the MRC Holland P008 PMS2 kit. Briefly, probes for the target region are hybridised to the DNA sample, and upon successful hybridisation, are ligated together. The ligated oligonucleotide is amplified and analysed by capillary electrophoresis to assess copy number. MLPA for this patient supported by a relative reduction in peak height and presence of a deletion (due to there being one *PMS2* copy and two *PMS2CL* copies of exon 15 the dosage quotient was reduced to ~0.75 rather than ~0.5). Given the clinical history and range of molecular testing the variant was classified to be likely pathogenic. Nevertheless, given its location in exon 15, it was not possible to exclude the possibility of it occurring in *PMS2CL* and it remained incompletely characterised. To identify the deletion breakpoint and confirm the sequence was deleted from within *PMS2*, linked-read sequencing was performed. A linked-read molecule of approximately 24 kb (Figure 4.1) was required to characterise the exon 15 deletion. The performance metrics for this sample revealed the library had a short average molecule length of 10.7 kb. Consequently, there were insufficient reads extending from the gene-specific region to exon 15 to characterise the apparent deletion (Figure 4.10). Further assessment of reads that had mapped to the 3' end of *PMS2* revealed they correlated with RepeatMasker defined SINE elements (Figure 4.11). These sequence reads were interpreted to be mis-mapped alignments from sequences that co-occur within the *PMS2* exon 6-8 gene-anchored region.

4.2.8.3 Sample 5

The patient was referred for molecular diagnostic investigation of the Lynch syndrome genes (*MSH2*, *MSH6*, *MLH1* and *PMS2*) following a microsatellite instability high result (MSI-H) and normal immunohistochemistry for MLH1 and MSH2. Targeted hybridisation capture and short-read sequencing identified an apparent 2-bp heterozygous deletion in *PMS2* exon 13, c.2186_2187del p.(Leu729fs), which was predicted to result in a frameshift in the translated protein (Figure 4.12A). While the variant had been previously reported as a pathogenic sequence change in the literature, having been identified in patients with constitutional mismatch repair deficiency syndrome, the variant has also been reported at significant frequencies in control datasets. Given that *PMS2* exon 13 has high sequence homology to *PMS2CL* it was not possible to exclude the variant from the pseudogene in standard short read data. By selecting only the BX-tags (the molecular identifier of reads originating from the same DNA fragment) which were aligned to the gene-specific region, we were able to demonstrate that linked reads anchored in the PMS2 gene did not contain the deletion, supporting that the deletion was in fact located in *PMS2CL* (Figure 4.12B). This observation reduces the clinical risk of PMS2-associated cancer in this patient.

4.2.8.4 Sample 6

Familiar testing by the diagnostic lab confirmed the patient to be heterozygous for the exon 13 pathogenic *PMS2* familial mutation, c.2192_2196del, p.(Leu731fs). The mutation was identified by Sanger sequencing, so a direct comparison to hybridisation capture short read sequencing data was not possible. Nevertheless, linked-read analysis verified the variant was specific to *PMS2* rather than *PMS2CL* (Figure 4.13). The data demonstrate how the selection of sequence reads from gene-anchored molecules (using the BX-tag) generates the most specific aligned-read dataset; for *PMS2CL* alignments there was a reduction in mis-mapped deletion-containing reads from 16 (when considering native linked-read alignments) to 0 (when only considering BX-tag selected reads).

4.2.8.5 Sample 7

A male patient affected with prostate and breast cancer was referred for analysis of 15 hereditary cancer predisposition genes, including *PMS2*. Molecular diagnostic testing did not identify any pathogenic variants. Nevertheless, sequence reads supporting the likely benign pseudogene derived *PMS2* variant c.2324A>G were detected. The variant was detected with an apparently low variant allele fraction (0.145), which was possibly due to the *PMS2* sense strand encoded non-reference nucleotide Chr7:6017340C corresponding to the antisense encoded *PMS2CL* wild type nucleotide Chr7:6786715G. This prompted assessment of this variant in the linked-read dataset. Both native linked-reads alignments and those sequence reads selected from molecules anchored to unique gene-specific region were considered. Overall, linked-read data had a variant allele fraction that most closely resembled a 50:50 allelic split. This was 0.361 (native alignments) and 0.268 (BX-tag selected alignments) compared to 0.145 (standard-of-care non-linked read dataset) (Figure 4.14). Considering the findings from previously described linked-read libraries, it was surprising that the native read alignments generated an allele fraction that was more consistent with a 50:50 allelic split (0.361) compared to the BX-tag selected alignments (0.268). This might be due to the variant occurring in *PMS2* exon 14 (compared to exon 13 for Samples 5 and 6) and the correspondingly lower read depth that was achieved (a cumulative total of 41).

4.2.8.6 Sample 8

A female patient affected with endometrial cancer was referred for analysis of pathogenic variants in Lynch syndrome associated genes. Immunohistochemical analysis had previously revealed a loss of MLH1 and PMS2 staining. No coding or splice-site variants were identified in *MSH2*, *MSH6* or *MLH1* by targeted capture hybridisation and short-read sequencing. Analysis of *PMS2* by targeted hybridisation and short read sequencing identified reads supporting the variant c.2350G>A, p.(Asp784Asn). This missense change results in a conservative amino acid substitution of a moderately conserved residue. While the variant has been previously reported in population control datasets, frequency estimates may be unreliable due to the high sequence homology of the locus. Molecular diagnostic classification determined the variant to be of uncertain clinical significance.

Linked-read sequencing was performed to assess whether the 0.32 variant allele fraction obtained from the standard-of-care hybridisation capture dataset could be improved to more closely resemble 0.50 (the expected value for a heterozygous germline variant). Analysis of reads derived from molecules extending into the unique region generated an allele fraction of 0.38, with no variant-containing reads (Chr7:6786741A) being observed at the *PMS2CL* locus (Figure 4.15). A lower allele fraction (0.28) was observed when the native linked-read alignments were reviewed this is likely due to a proportion of reads being absorbed to the *PMS2CL* locus (Chr76786741A: 3.6%). While these data have not changed the interpretation of the identified variant, they support the view that the linked-read workflow has the potential to generate more specific, and therefore accurate alignments.

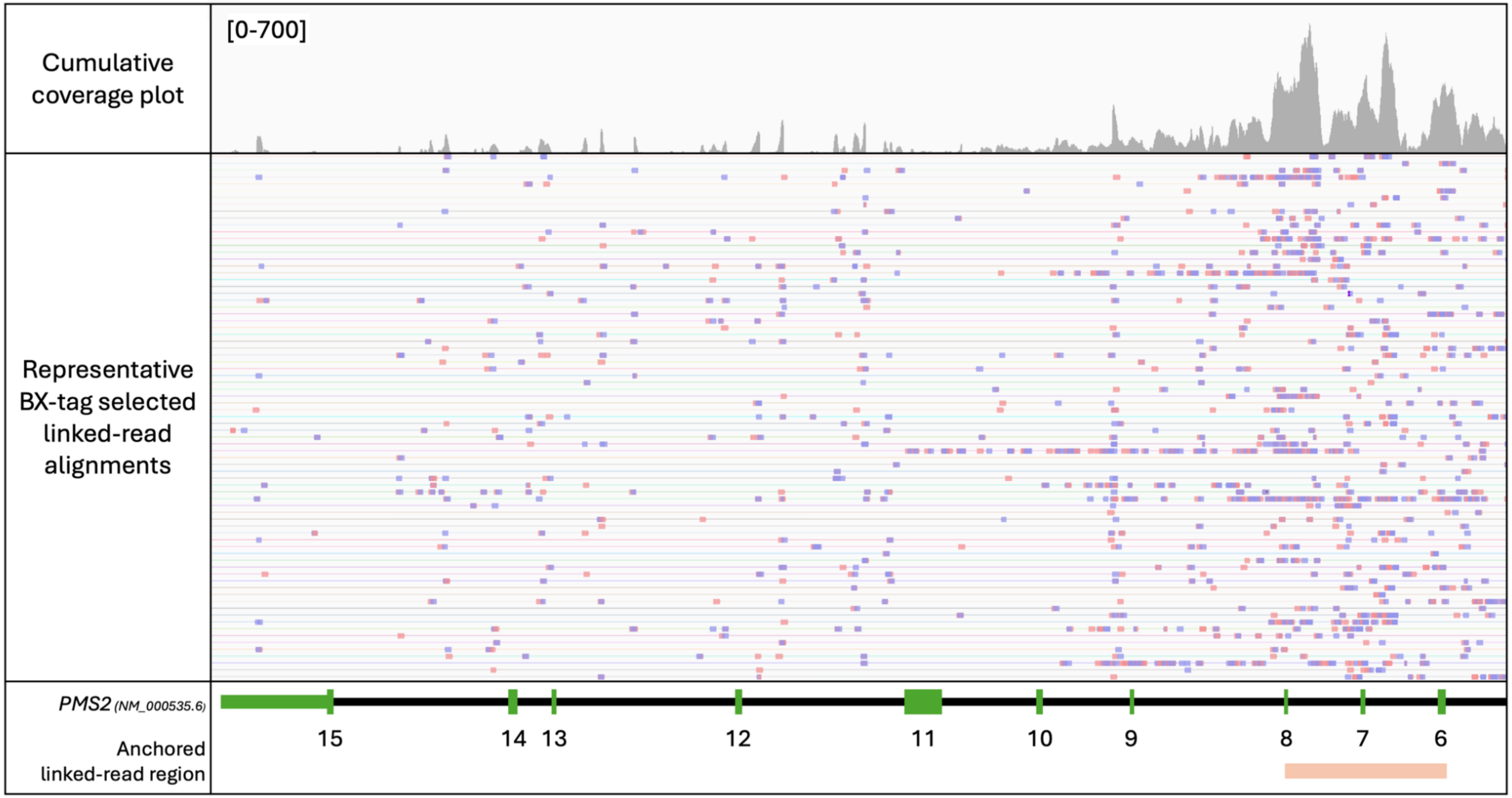

Figure 4.10: Linked-read molecules for Sample 4 that were anchored in PMS2 exons 6-8. An insufficient number, of sufficiently long molecules were identified to enable characterisation of the apparent heterozygous *PMS2* exon 15 deletion. The source molecule of aligned reads is determined by a unique molecular index.

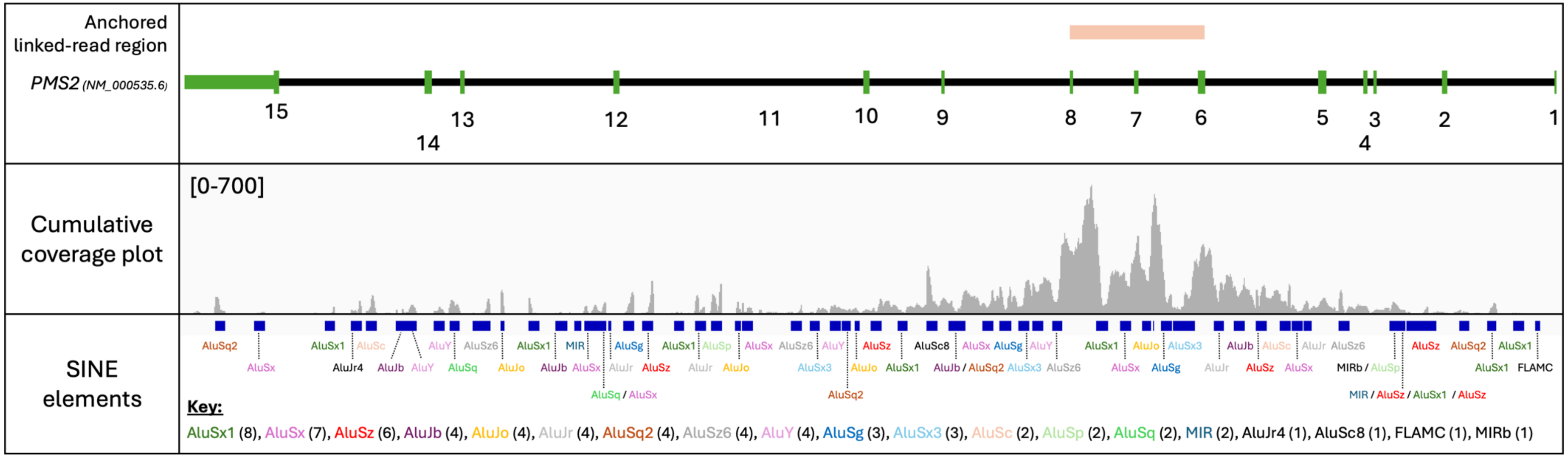

Figure 4.11: Sample 4 sequence read alignments that map out with the anchored linked-read interval predominantly map to low complexity SINE elements repeats that are represented within the unique sequence. It is likely the reads have been mis-mapped due to their homology to low-complexity sequences located within exons 6-8. Annotated SINE elements are coloured coded when observed twice or more (see key for the number of instances of each element).

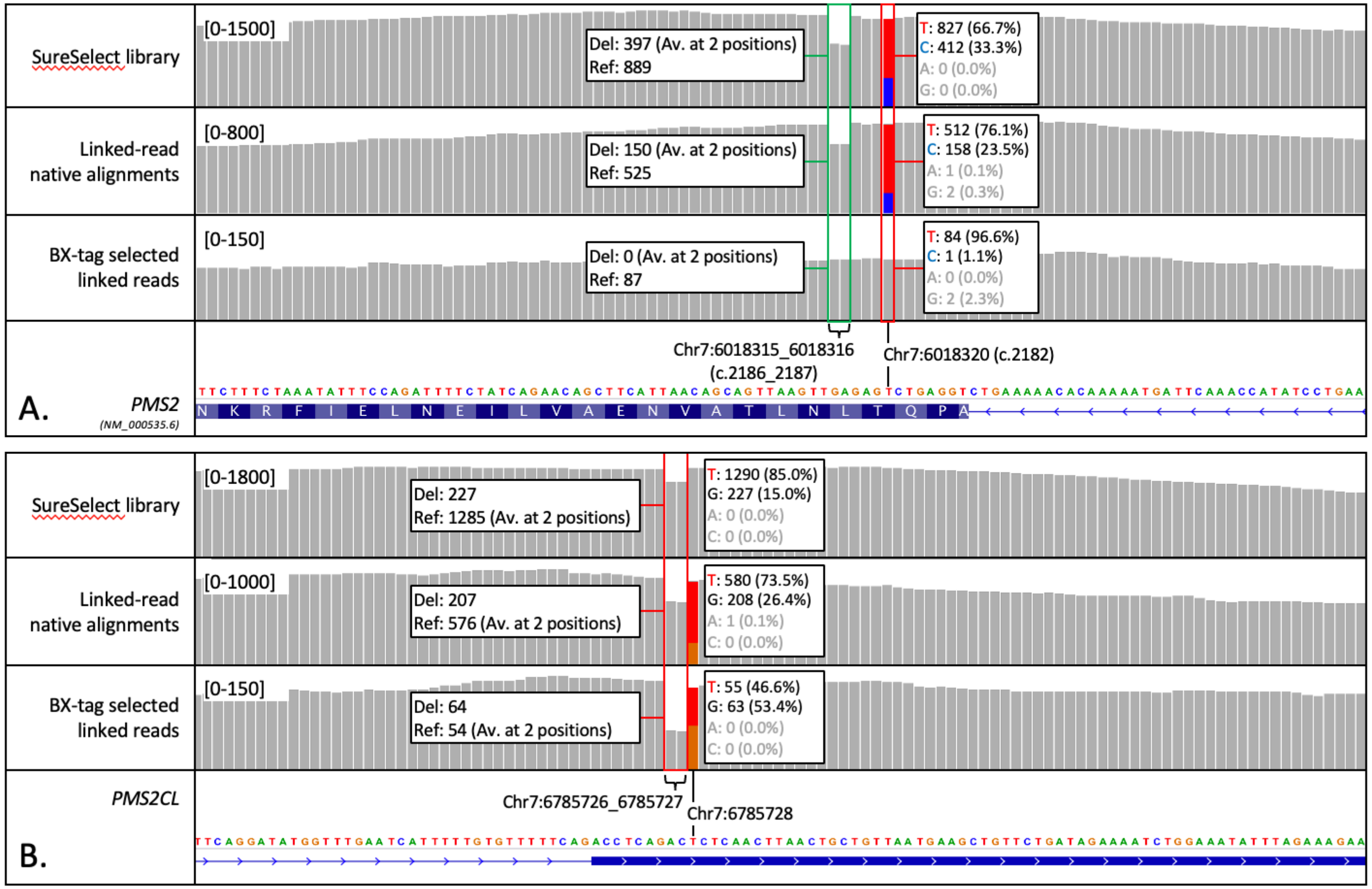

Figure 4.12: IGV cumulative coverage plots for Sample 5 at homologous regions of the *PMS2* and *PMS2CL* genes. (A) The SureSelect library track is the standard-of-care hybridisation capture and short-read NGS data. Note the difference between native linked-read alignments and those sequence read alignments selected from gene-anchored molecules defined by the BX-tag **(B)**. The putative *PMS2* c.2186_2187del variant is absent from BX-selected alignments, which map specifically to *PMS2CL*. *Y*-axis labels for each track denote the depth of coverage. As expected, the depth-of-coverage is lower for BX-tagged selected reads (which are required to be from a gene-anchored molecule).

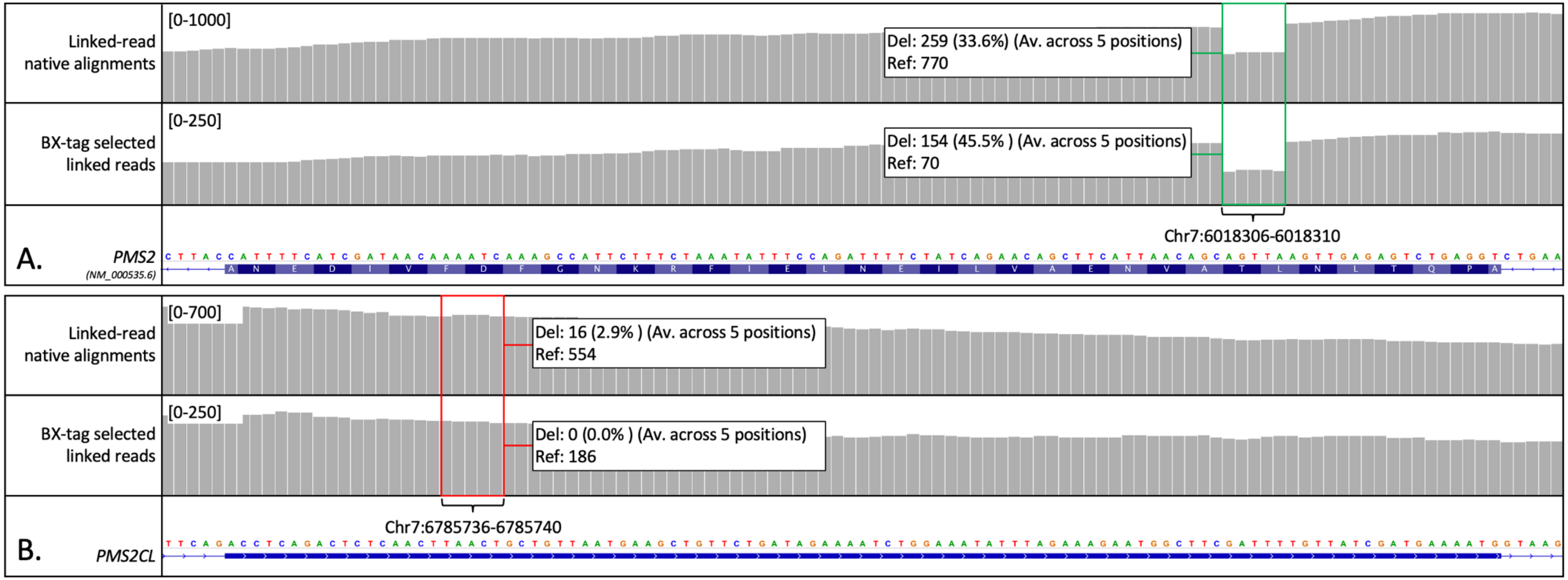

Figure 4.13: IGV cumulative coverage plots for Sample 6 at homologous regions of the *PMS2* and *PMS2CL* genes. (A) The heterozygous 5-bp deletion, c.2192_2196del, is visible in *PMS2* exon 13 (highlighted green). When selecting reads that are derived from molecules located in the unique region the total number of reads is reduced, and the allelic balance better resembles the 50:50 expected ratio (the proportion of deletion-specific reads increases from 33.6% to 45.5%). **(B)** Alignments at *PMS2CL* show a reduction in deletion-containing reads (from 16 to 0) when considering only those from molecules mapped to the unique region (highlighted red). *Y*-axis labels for each track denote the depth of coverage.

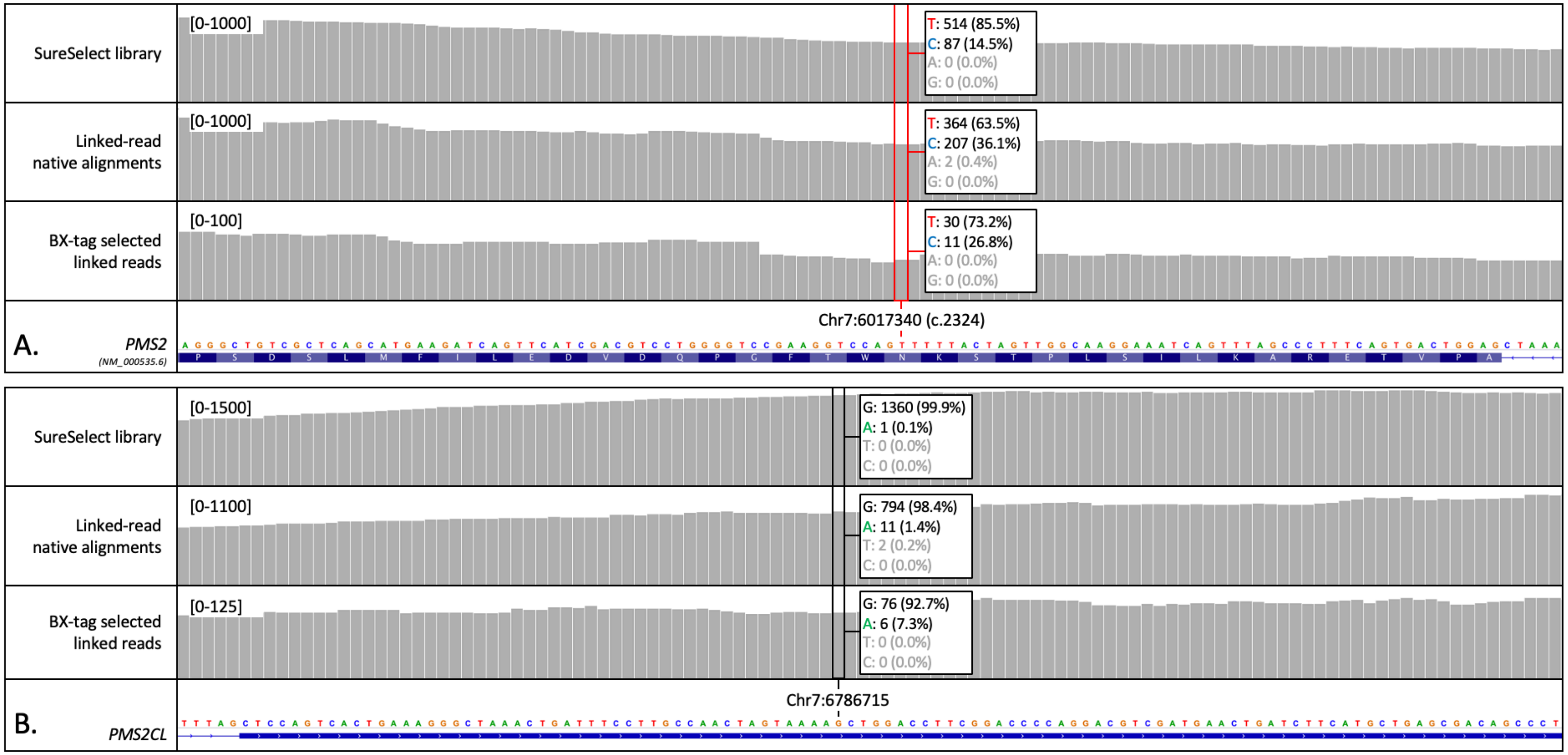

Figure 4.14: IGV cumulative coverage plots for Sample 7 at homologous regions of the *PMS2* and *PMS2CL* genes. (A) The *PMS2* non-reference nucleotide fraction at position Chr7:6017340 was observed to be 0.145 following standard-of-care hybridisation capture enrichment; the non-reference nucleotide matches the wild-type base in *PMS2CL* and likely resulting in a considerable number of mis-mapped alignments. Overall, linked-read alignments generated an allelic fraction that was closer to the expected 50:50 ratio. Although it was the non-selected, native aligned linked-read track, that had the highest variant allele fraction (0.361), the cumulative read depth for BX-tag selected reads was low (41×). **(B)** While there is some variability in the proportion of non-reference reads at the corresponding *PMS2CL* variant position, the absolute number of non-reference reads was low. Note that the *PMS2CL* reference nucleotide at position Chr7:6786715 corresponds to the non-reference *PMS2* nucleotide.

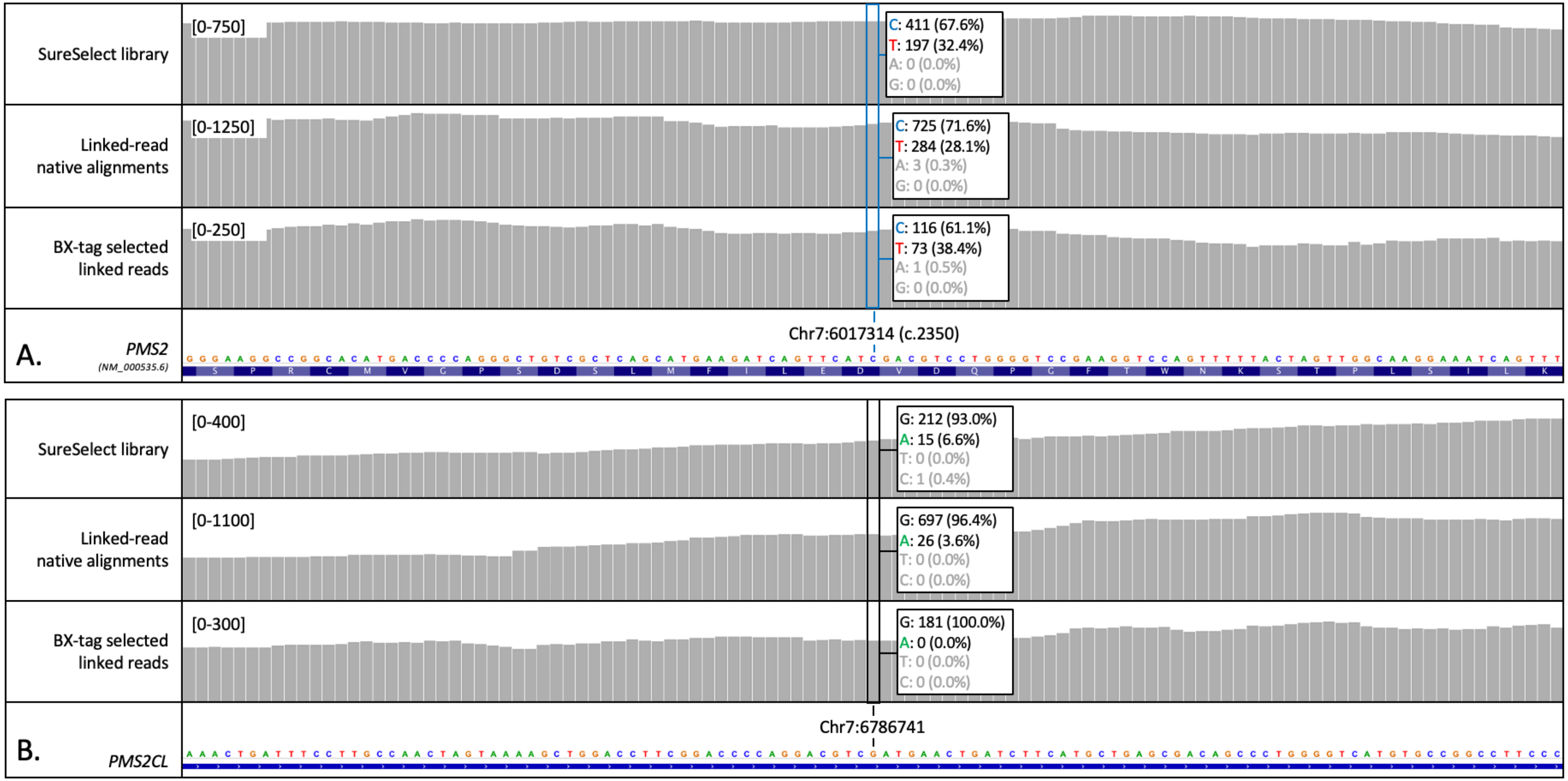

Figure 4.15: IGV cumulative coverage plots for Sample 8 at homologous regions of the *PMS2* and *PMS2CL* genes. (A) The variant allele fraction at position Chr7:6017314 was observed to be 0.324 following standard-of-care hybridisation enrichment. The was marginally improved to 0.384 following the selection of sequence reads that were from molecules in a gene-anchored region. **(B)** There were a notable absence of non-reference sequence reads at the homologous Chr7:6786741 *PMS2CL* locus the gene anchored linked-read dataset.

4.2.9 Verification of *PMS2*-specific genotypes by LR-PCR supports linked-read data.

For samples in which the interrogated variants are located within the genomic region targeted by the Edinburgh amplicon, long-range PCR was performed on the patient DNA to verify the linked-read data findings. PCR products were fragmented and short-read sequencing was performed. Linked-read observations were verified for each of Samples 5, 6, 7 and 8 (Figure 4.15). For Sample 7 the variant allele fraction (for the non-reference base) was 0.18; this non 50:50 allelic balance was likely due to sequence reads from the pseudogene-derived variant, c.2324A>G (chr7(GRCh37):g.chr7:6017340T>C), being mapped to *PMS2CL*. (None of the target variants were located within the smaller Set 3 amplicon.)

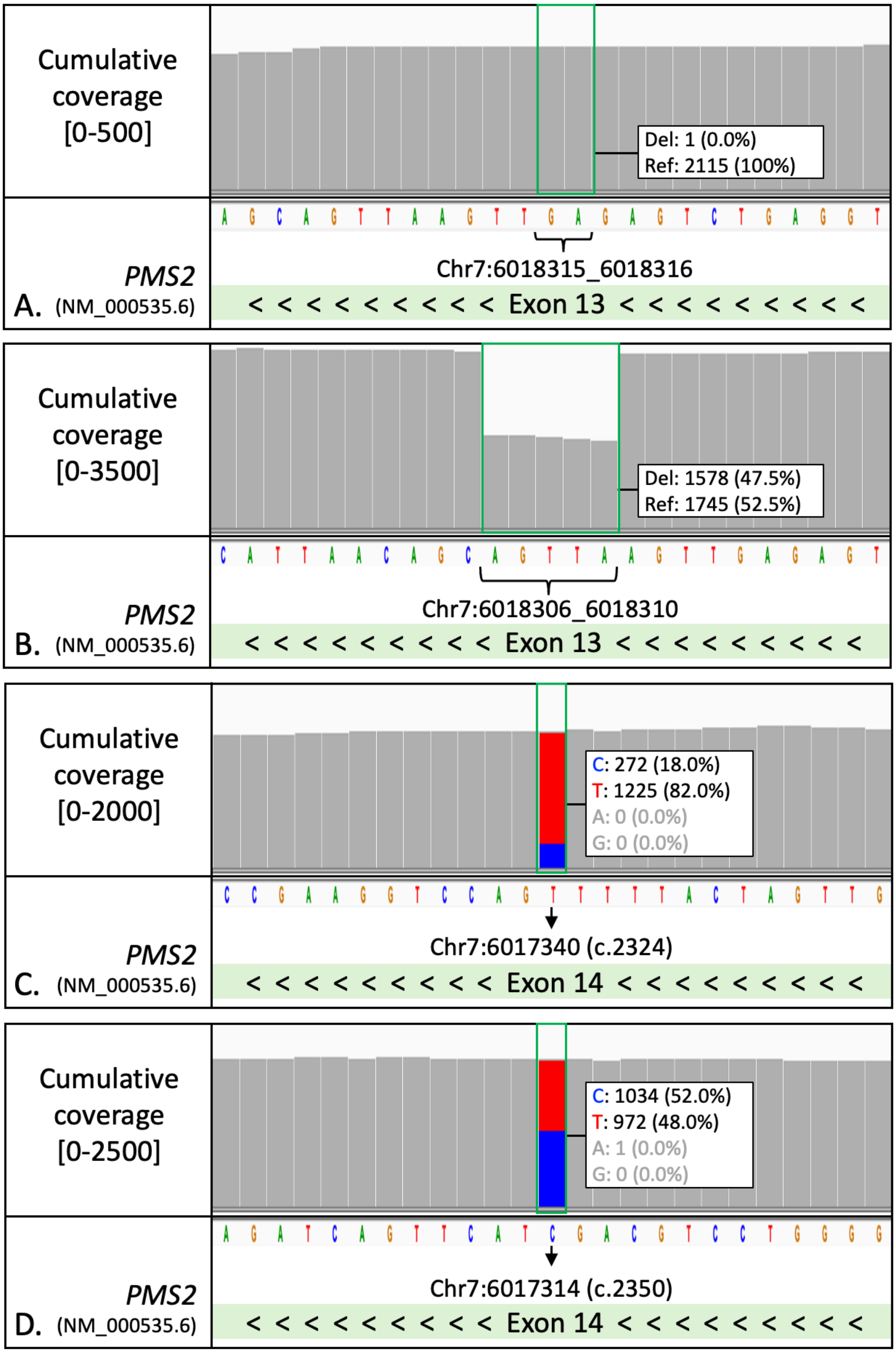

Figure 4.16: Verification of interrogated *PMS2* variants using the LR-PCR Edinburgh assay and short-read sequencing. (A) Sample 5. **(B)** Sample 6. **(C)** Sample 7. **(D)** Sample 8. The *y*-axis cumulative coverage scale is defined below the track title. *PMS2* is encoded on the antisense strand; arrows indicate the direction of transcription.

4.2.10 Retrospective assessment of the fragment profiles of genomic DNA

When research into the performance of linked-read technology was originally performed, the ability to assess the DNA fragment distributions was limited to pulse-field gel electrophoresis. The large mass of DNA required, combined with the limited availability of clinical specimens, prohibited this analysis. More recently the laboratory has bought and commissioned a Femto Pulse (Agilent Technologies) which permits PFGE of nanogram quantities of starting material. The eight genomic DNA samples used to prepare linked-read libraries were assessed by the Femto Pulse, with the corresponding traces displayed in Figure 4.16. Consistent with a short linked-read molecule length, and comparatively poor library metrics, the DNA profile for Sample 4 showed a bimodal fragment distribution with a significant proportion of short, degraded fragments. Although Sample 6 also displayed a bimodal distribution, the peak for larger molecular mass DNA was bigger than that which was observed for Sample 4 (33.8 kb versus 20.9 kb, respectively). The traces for all other samples (particularly Samples 1, 7 and 8) revealed intact high molecular weight DNA with no degradation.

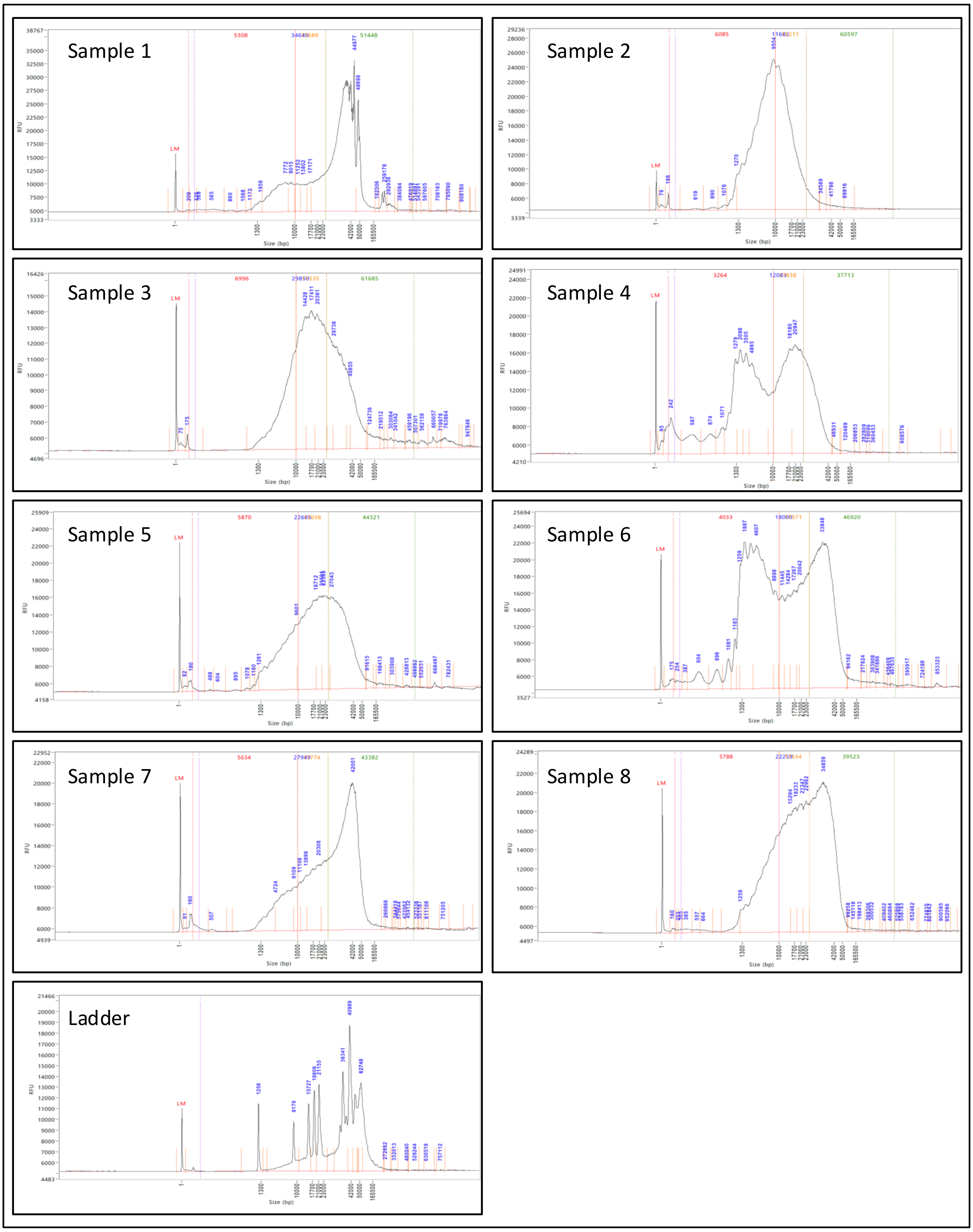

Figure 4.17: Femto Pulse traces for each of the eight genomic DNA samples investigated by linked-read sequencing. Note the bimodal peaks for Samples 4 and 6. *X*-axis displays the log-scaled fragment length in base pairs, with peak sizes determined from markers of known size run in the ladder lane.

4.3 Discussion

The work presented in this chapter sought to develop a linked-read sequencing assay for the analysis of *PMS2* and its pseudogene *PMS2CL*. It was anticipated that this would overcome a significant limitation of standard short-read sequencing target enrichment workflows which prohibit the unambiguous alignment of sequence reads at homologous loci. The sensitivity of the assay, for variant detection, was benchmarked using the comprehensively characterised GIAB sample (NA12878), before a series of clinical cases (with known or incompletely resolved *PMS2*/*PMS2CL* variants) were next analysed.

For each sequenced library, assay performance metrics were first assessed. These data demonstrated that the pooling of individual libraries generated an approximately even distribution of sequence reads, and the majority of these reads (>96%) could be mapped to the human reference sequence. Nevertheless, for each library an unexpectedly high proportion of reads were identified as PCR duplicates. Read pairs are classed as “PCR duplicates” when they have the same mapping coordinates as one or more other read pairs from the same sequencing library. The most likely explanation for this scenario is that the sequenced fragment has arisen due to PCR amplification of the original molecule, rather than it being a separate biological observation. As PCR duplicate reads do not add additional support for the identification of reference or non-reference bases, they are typically filtered out of the dataset at the analysis stage. PCR duplicate rates are influenced by the amount of starting DNA, the number of rounds of thermocycling performed, as well as the configuration and number of hybridisation probes used. For a typical hybridisation capture experiment, the proportion of duplicate reads is typically 2-8% (Marosy, 2018).It is notable that a previous study, which used linked-read whole genome sequencing in combination with exome-targeted hybridisation probes manufactured by Agilent Technologies, reported duplicate rates of 0.7-6.45% (Zheng *et al.,* 2016). High duplicate rates impact the total cost of a workflow as a higher yielding sequencing cartridge is required to obtain a sufficient depth-of-coverage. One factor that may have contributed to the high duplicate rate is the low DNA mass requirement (5 ng) for the 10X Genomics whole genome library preparation; a low starting mass reduces the number of genome copies in the specimen. The impact of a low starting mass of DNA was likely exacerbated by the requirement to use 187.5 ng of prepared library for the hybridisation capture step. Additionally, GEM barcoding involves PCR amplification which introduces a third PCR step into the workflow (a process that is not required for standard hybridisation capture workflows). Future optimisation of the assay would focus on a reduction in the number of rounds of amplification and a lowering of the mass of prepared library used for targeted hybridisation capture.

A further observation following assessment of the assay performance metrics is the low proportion of reads mapping to the target region. For a typical hybridisation capture experiment the on-target rate (the number of reads mapping to regions of interest) is expected to be >80%. By contrast, on-target rates for the custom linked-read workflow were approximately 15%. An increased proportion of off-target reads is typically due to a reduction in the performance of the hybridisation capture reagent (rather than probes binding non-specifically to regions that were not targeted). For the custom reagent developed for this project, regions of low sequence complexity (those present in the RepeatMasker track), were excluded from the design and no probes were synthesised to the corresponding regions within the target interval. Visual inspection of ad-hoc genomic regions revealed a consistently sparse uniform distribution of sequence reads (rather than specific clusters of high sequence depth) supporting the view that the capture was inefficient, rather than the probes being bound to specific off-target regions. While manufacturers invest considerable time and money into developing robust and performant assays, the presented data was generated using linked-read library preparation reagents from 10X Genomics and hybridisation capture probes and wash reagents from Twist Biosciences. For effective oligonucleotide hybridisation, it is recognised that the salt concentration is a critical variable; one hypothesis for the lower-than-expected on target rate is that the 10X Genomics prepared libraries were not at an optimum molarity. The previously cited study by Zheng *et al.,* (2016) (linked-read whole genome library preparation in combination with exome sequencing) achieved on-target rates of between 52-62%

(Zheng *et al.,* 2016) using probes manufactured by Agilent Technologies; a future assessment of probes from this manufacturer has been considered. Similar to the high duplicate rate, the observation that a lower-than-expected proportion of reads mapped to the target interval impacts the assay by requiring the use of a higher yielding cartridge; such cartridges are correspondingly more expensive.

While the predominant aim of this work was to assess the utility of linked-read sequencing at homologous loci it was first necessary to establish a bioinformatics pipeline. To evaluate the approach, known genotypes from the GIAB sample were first evaluated. The identification of SNVs from non-homologous (high confidence) regions was highly sensitive (97.83%). False negative calls were either identifiable from the raw sequence reads but had been filtered out due to aggressively applied bioinformatics parameters, or there were insufficient reads covering the variant location. The assessment of insertion-deletion variants produced a sensitivity value of 80.37%, which is notably lower than for SNVs. The majority of false negative indel calls were due to variants occurring at the proximity of poly(N) tracts, genomic features that are notoriously challenging to characterise by short read sequencing (Fang *et al.,* 2014). Nevertheless, these data support the potential wider use of linked-read library preparation beyond its use for the assessment of homologous gene regions. Targeted panel sequencing, where several genes are concurrently sequenced, is the standard approach to molecular genetic diagnosis. It could therefore be envisioned that linked-read library preparation, for analysis of the *PMS2* locus, is used in combination with targeted capture of additional hereditary cancer predisposition genes.

For one of six clinical samples sequenced, the target variant could not be characterised. This case (Sample 4) had an apparent heterozygous deletion of exon 15, the *PMS2* final exon. The case was sequenced in an effort to characterise the variant at single nucleotide resolution. Cumulative read depth, beyond the anchored unique sequence (exons 6-8) was low, and interrogation of the reads that had mapped to *PMS2* beyond exon 8 revealed that they were mostly located within low complexity SINE elements that were represented within the exon 6-8 region, so they were likely to be apparently miss-aligned reads (Figure 4.10). Retrospective analysis of the fragment profile of this sample was performed using a Femto Pulse that had been commissioned after this work was performed. This revealed a bimodal size distribution with the low molecular weight peak having a high molarity of degraded fragments. This case emphasises the requirement to use high molecular weight fragments. Prior to the availability of the Femto Pulse, fragment profile assessments were limited to pulse-field gel electrophoresis that is both a time-consuming and temperamental technique that only provides a qualitative (rather than quantitative) interpretation. Importantly, the quantity of DNA required for PFGE is significant, and frequently prohibitive for clinical cases with scarce availability of specimen. By contrast, the Femto Pulse is ultrasensitive and can analyse input concentrations as low as 5 pg/µl.

Clinical laboratories typically receive specimens from one or more peripheral centres; the adoption of novel technologies into routine clinical practise therefore requires an assessment of different sample types. In this regard, Sample 2 was a freshly obtained specimen extracted using the bead based Chemagic 360 instrument, the default pathway for extraction in the diagnostic laboratory. Assay performance metrics and the subsequent Femto Pulse trace demonstrated that high quality data could be obtained from this DNA sample, suggesting that the linked-read workflow could be adopted for routine clinical use.

The historical handling of samples adds further complexity to any interpretation of appropriate extraction pathways. Repeated freeze-thawing cycles of DNA samples impacts their fragment distribution, resulting in increased degradation, with clinical samples often subjected to multiple tests which results in repeated freeze thaw cycles. While an auditable log of a sample’s provenance would be ideal, this is seldom available in the case of samples received from external laboratories, and a full understanding of workflow performance requires an empirical assessment of real-world activities. An ongoing assessment of the assay’s performance will therefore be critical to understanding the limitations of different DNA extraction reagents.

An important part of the work described here is ascertaining the ability of this linked-read technology to provide a definitive variant classification in scenarios where the current standard of care is unable to do so, due to technical limitations of the existing test. In sample 3 the utility of linked reads is demonstrated to great effect. The whole gene deletion of PMS2 was estimated to be 88 kb in size based on shallow whole genome sequencing, but the precise location of the breakpoints was not determined due to the ambiguous mapping of reads enriched by the existing hybrid-capture workflow.

Initially the linked read data for sample 3 was sorted into haplotype groups in order to determine the parental origin of the deletion, allowing determination of the distal breakpoint in the CCZ1 gene. This data suggested the proximal breakpoint was in EIF2AK1; consequently, the linked reads were anchored in CCZ1 exon 1 allowing the mapping of the inverted region and demonstrating the breakpoint was indeed in this gene.

The methods deployed here in combination with the linked reads data could readily be applied to other clinical scenarios. As we have previously reported, phasing of variants using long-PCR approaches is challenging and can be hampered by the generation of chimeric amplicons (McClinton, Watson, *et al.,* 2023). Furthermore, whilst we have previously demonstrated alternative approaches to breakpoint mapping and characterisation of structural change using WGS (Watson *et al.,* 2014, 2016b), the data presented in this case suggests linked reads may be a suitable alternative in the targeted NGS arena.

While massively parallel sequencing technologies have expanded the testing repertoire available to patients with suspected hereditary cancer predisposition, the large gene panels used for this testing also identify a large number of variants of uncertain significance (VUS). A recent, large cohort study of patients referred for gene panel testing found that at least one VUS was detected in approximately 41% of the patients tested (E. Chen *et al.,* 2023). While the data in this study suggests only ~20% of VUS will later be reclassified to likely pathogenic or pathogenic, reporting of these variants does not always serve to alleviate patient anxiety and may in fact instigate excess clinical surveillance and unnecessary familial testing (Burke *et al.,* 2022). It is also established that the incidence of VUS is higher in some ethnicities due to under-representation in the population data sets on which variant classifications are based

(Ndugga-Kabuye & Issaka, 2019). This is pertinent in the case of sample 5, as the VUS reported in this patient was noted at the time to have a minor allele frequency of ~2% in African populations (Leongamornlert *et al.,* 2014). However, due to the uncertainty of whether the variant was in the gene or the pseudogene, it was still reported as class 3 variant of uncertain significance, leading to an inconclusive outcome for the patient. The linked-read data confirmed that the variant was in fact located in *PMS2CL*, reducing the risk of *PMS2* associated cancer in this patient. Since this work was carried out, there have been further reports of the variant being observed at high frequency in non-European populations (Segura *et al.,* 2024), which have also located the variant in PMS2CL, highlighting the importance of ethnic diversity in data sets in order to facilitate variant classification, and the value of endeavours such as the Pangenome Project to improve the diversity of reference genomes. Nevertheless, there are still rare incidences where common PMS2CL variants, including this one, can also be detected in PMS2 (Pan *et al.,* 2019.), demonstrating the need for an NGS test which allows discrimination between the two loci. This case is therefore an important example of how linked reads could be applied to improve patient outcomes.

From a diagnostic point of view, the data here has demonstrated the ability of this assay to definitively characterise variants arising in PMS2 and PMS2CL where the current standard of care technique had been unable to do so. While there was potential to perform a larger validation study comparing the two methods in parallel, unfortunately shortly after this work was performed 10X Genomics announced that they would be discontinuing their manufacturing of linked-read reagents from June 2020. This revision of their product portfolio coincided with a period of litigation over several patents held by BioRad, a competitor organisation, that related to droplet-based methodologies and their use for genomic analysis.

Since that time, several alternative linked-read products have been commercialised. The most notable of these is Transposase Enzyme-Linked Long-read Sequencing, known as TELL-Seq^{TM} (Universal Sequencing Technologies, USA) (Z. Chen *et al.,* 2020); the workflow uses bead-based technology to barcode HMW DNA fragments. In comparison to the 10X Genomics workflow the method does not require any specialist laboratory equipment in order to perform the assay, therefore making it attractive to laboratories that have a limited budget for capital expenditure. Illumina have a linked-read comparable product, Complete Long Read Prep(Illumina, no date), through which HMW DNA is tagmented into multi-kilobase fragments and enzymatically "land-marked" at the single molecule level. The marked fragments are then amplified and used to create a standard short-read library. The short reads sequences are then linked bioinformatically based on the molecular landmarks that were previously generated. Illumina have also recently announced constellation mapping, whereby library preparation is completed on the flow cell surface directly followed by sequencing, enabling linkage of sequencing reads arising from the same molecules by their adjacent positions on the flow cell.

In addition to the techniques enabling retrospective bioinformatic compilation of long reads, the evolution of long read sequencing is such that native sequencing of gDNA to analyse genes contiguously is now a viable alternative. The widely reported use of adaptive sampling technology in order to enrich for targets from whole genome long read sequencing in real time is particularly attractive for diagnostic screening panels.

4.4 Summary

With this work, we have been able to demonstrate that linked-read libraries can be combined with a custom hybridisation capture method to successfully sequence a clinically relevant cancer associated gene and its most paralogous pseudogene, providing the capability to distinguish between the two regions in aligned sequence data and identify the origin of *PMS2* and *PMS2CL* variants. Of the six clinical cases included in the experiment, we have successfully characterised the variants in five of the samples. While the specific technology used here has since been discontinued, the prospect of alternative linked-read methods, as well as the emergence of long-read sequencing, there is great potential to develop alternative workflows for the analysis of clinically relevant genes where the presence of pseudogenes may hamper diagnosis by standard short-read sequencing technologies.

5. Investigating long read nanopore sequencing to improve the molecular diagnosis of rare disease

5.1 Introduction

Short-read sequencing-by-synthesis instruments developed and manufactured by Illumina have become the most widely adopted platforms for the identification of pathogenic variants causing rare disease. Nevertheless, the sensitivity of these workflows remains incomplete, with an unknown proportion of disease-causing variants refractory to detection. This is typically either due to the chemistry being unable to generate sequence data for a challenging region of the genome (*e.g.* a locus with a high GC content), or because the sequence reads that are generated are too short to be unambiguously aligned to the reference genome. Furthermore, while some diagnostic referrals are now being processed by short-read whole genome sequencing a significant proportion of cases are still analysed by targeted hybridisation capture enrichment of the coding exons and their immediate splice junctions. In this scenario variants may be refractory to detection by virtue of not being located within genomic regions targeted by the custom designed target enrichment reagent.

Now, third-generation whole genome sequencing platforms, such as the PromethION (an instrument developed by Oxford Nanopore Technologies), have the potential to identify variants that are refractory to detection by existing standard of care approaches. To understand the strengths and weaknesses of these new platforms it is necessary to have both first-hand experience of the workflow and identify real-world clinical questions that can be assessed. In this regard, one cohort of patients that are of particular interest to the molecular diagnostic community are those cases in whom a single bona-fide heterozygous pathogenic variant has been identified in an autosomal recessive disease gene that matches the disease phenotype. These patients may either be incidental carriers of the single heterozygous pathogenic variant, or there may be a missing “second hit” pathogenic variant that was refractory to detection by the initial analysis.

5.2 Context of the investigation

In 2023 the laboratory was funded as a Genomic England “Pathfinder” site as part of the Cancer 2.0 whole genome long-read sequencing project (https://www.genomicsengland.co.uk/initiatives/cancer). The purpose of the funding was to install and commission a PromethION nanopore sequencer to establish a long-read workflow for tumour and matched germline sequencing (working to diagnostic standards throughout). As part of the project, I led the technical implementation of the workflow for the North-East and Yorkshire Genomics Laboratory Hub (https://ney-genomics.org.uk/). Following sequencing, raw data (POD5 files) was transferred to a secure data centre for processing and interpretation by Genomics England scientists. Tumour and germline specimens underwent a pairwise comparison with their corresponding standard-of-care short-read datasets; variants that were identified or missed by each platform were evaluated to assess the strengths and limitations of the respective workflows.

In parallel with the Cancer 2.0 activity, we sought to understand the utility of the infrastructure in the context of inherited rare recessive disease; this chapter reports these data. Two cases (from different families) with single heterozygous mutations, identified in genes associated with autosomal recessive disease, were identified as a result of standard-of-care testing. The presenting phenotypes in these cases, matched those reported in the literature, suggesting it was highly likely that a “second hit” variant remained undetected following short read sequencing.

Case 1 was a patient referred for analysis of intestinal failure and congenital diarrhoea genes, and in whom standard-of-care analysis identified a heterozygous pathogenic variant in the autosomal recessive disease gene *TTC37*.

Case 2 was a pregnancy for which ultrasound screening had identified encephalocele and cystic kidneys in the developing fetus. The family was referred for prenatal trio exome sequencing which identified a maternally inherited pathogenic variant in the Meckel-Gruber associated gene *TMEM67*.

5.3 Case 1: Results

5.3.1 Standard-of-care investigation

A female patient was referred for genetic testing as an infant, with intrauterine growth restriction (IUGR), bilateral dislocation of hips, cleft palate, and hypoglycaemia. SNP array analysis did not detect any significant imbalances. Subsequently, at aged 1 year the patient was referred for targeted next generation sequencing analysis due to the presentation of woolly, brittle hair, and intestinal failure with congenital diarrhoea, that required parenteral nutrition. Hybridisation capture and short-read sequencing was performed for the 45 genes on the intestinal failure or congenital diarrhoea gene panel (R331 v.2.0) defined by PanelApp (Stark *et al.,* 2021) (Table 5.1).

Table 5.1: Summary of genes that are screened as part of the R331 gene panel.

HGNC Gene name	Location	Strand	MANE Select transcript	Genomic coordinates*	Number of exons
ADAM17	2p25.1	-	NM_003183.6	Chr2:9488486-9555830	19
ADAMTS3	4q13.3	-	NM_014243.3	Chr4:72280969-72569221	22
ANGPTL3	1p31.3	+	NM_014495.4	Chr1:62597520-62606313	7
AP1S1	7q22.1	+	NM_001283.5	Chr7:101154476-101161276	5
APOB	2p24.1	-	NM_000384.3	Chr2:21001429-21044073	29
ARX	Xp21.3	-	NM_139058.3	ChrX:25003694-25015965	5
CCBE1	18q21.32	-	NM_133459.4	Chr18:59430939-59697423	11
CD55	1q32.2	+	NM_000574.5	Chr1:207321678-207360966	10
CLMP	11q24.1	-	NM_024769.5	Chr11:123069872-123195248	7
CTLA4	2q33.2	+	NM_005214.5	Chr2:203867771-203873965	4
DGAT1	8q24.3	-	NM_012079.6	Chr8:144314584-144326852	17
EGFR	7p11.2	+	NM_005228.5	Chr7:55019017-55211628	28
EPCAM	2p21	+	NM_002354.3	Chr2:47369311-47387020	9
FAT4	4q28.1	+	NM_001291303.3	Chr4:125314955-125492932	18
FLNA	Xq28	-	NM_001110556.2	ChrX:154348531-154374634	48
FOXP3	Xp11.23	-	NM_014009.4	ChrX:49250438-49264710	12
GUCY2C	12p12.3	-	NM_004963.4	Chr12:14612632-14696599	27
ICOS	2q33.2	+	NM_012092.4	Chr2:203936763-203961577	5
KMT2D	12q13.12	-	NM_003482.4	Chr12:49018978-49060794	55
LCT	2q21.3	-	NM_002299.4	Chr2:135787850-135837184	17
LRBA	4q31.3	-	NM_001364905.1	Chr4:150264435-151015284	57
MTTP	4q23	+	NM_001386140.1	Chr4:99574824-99623997	18
MYO5B	18q21.1	-	NM_001080467.3	Chr18:49822789-50195147	40
NEUROG3	10q22.1	-	NM_020999.4	Chr10:69571698-69573422	2
PCSK1	5q15	-	NM_000439.5	Chr5:96390333-96433248	14
PLVAP	19p13.11	-	NM_031310.3	Chr19:17351455-17377342	6
RFX6	6q22.1	+	NM_173560.4	Chr6:116877242-116932161	19
SAR1B	5q31.1	-	NM_016103.4	Chr5:134601149-134632828	7

SI	3q26.1	-	NM_001041.4	Chr3:164978898-165078496	48
SKIC2	6p21.33	+	NM_006929.5	Chr6:31959175-31969751	28
SLC10A2	13q33.1	-	NM_000452.3	Chr13:103043998-103066417	6
SLC26A3	7q22.3-q31.1	-	NM_000111.3	Chr7:107765469-107803223	21
SLC39A4	8q24.3	-	NM_130849.4	Chr8:144412414-144416844	12
SLC5A1	22q12.3	+	NM_000343.4	Chr22:32043261-32113029	15
SLC9A3	5p15.33	-	NM_004174.4	Chr5:470456-524449	17
SPINT2	19q13.2	+	NM_021102.4	Chr19:38264573-38292615	7
STX3	11q12.1	+	NM_004177.5	Chr11:59755376-59805878	11
STXBP2	19p13.2	+	NM_006949.4	Chr19:7637110-7647873	19
TERT	5p15.33	-	NM_198253.3	Chr5:1253167-1295068	16
TMPRSS15	21q21.1	-	NM_002772.3	Chr21:18269116-18403785	25
SKIC3 **(*TTC37*)**	5q15	-	NM_014639.4	Chr5:95463894-95554977	43
TTC7A	2p21	+	NM_020458.4	Chr2:46941224-47076123	20
WNT2B	1p13.2	+	NM_024494.3	Chr1:112508965-112530165	5
XIAP	Xq25	+	NM_001167.4	ChrX:123860053-123913972	7

*Genomic coordinates provided according to human reference genome build GRCh38 using the MANE Select transcript. (+): Sense strand. (-): Antisense strand.

The sequencing and data analysis was performed by the North East and Yorkshire Genomics Laboratory Hub using a custom hybridisation panel, leading to the identification of a single heterozygous variant of interest: c.2808G>A (NM_014639.4) in exon 28 of *TTC37* (Figure 5.1). The variant was predicted to create a premature termination codon in the translated protein p.(Trp936*). Despite a cumulative read depth of ≥30 reads across the coding sequence and canonical splice sites (+/- 2 bp) no deleterious second variant was identified. Additional comparative read-depth analysis was performed and did not identify a dosage change (exon gain or loss) within *TTC37*.

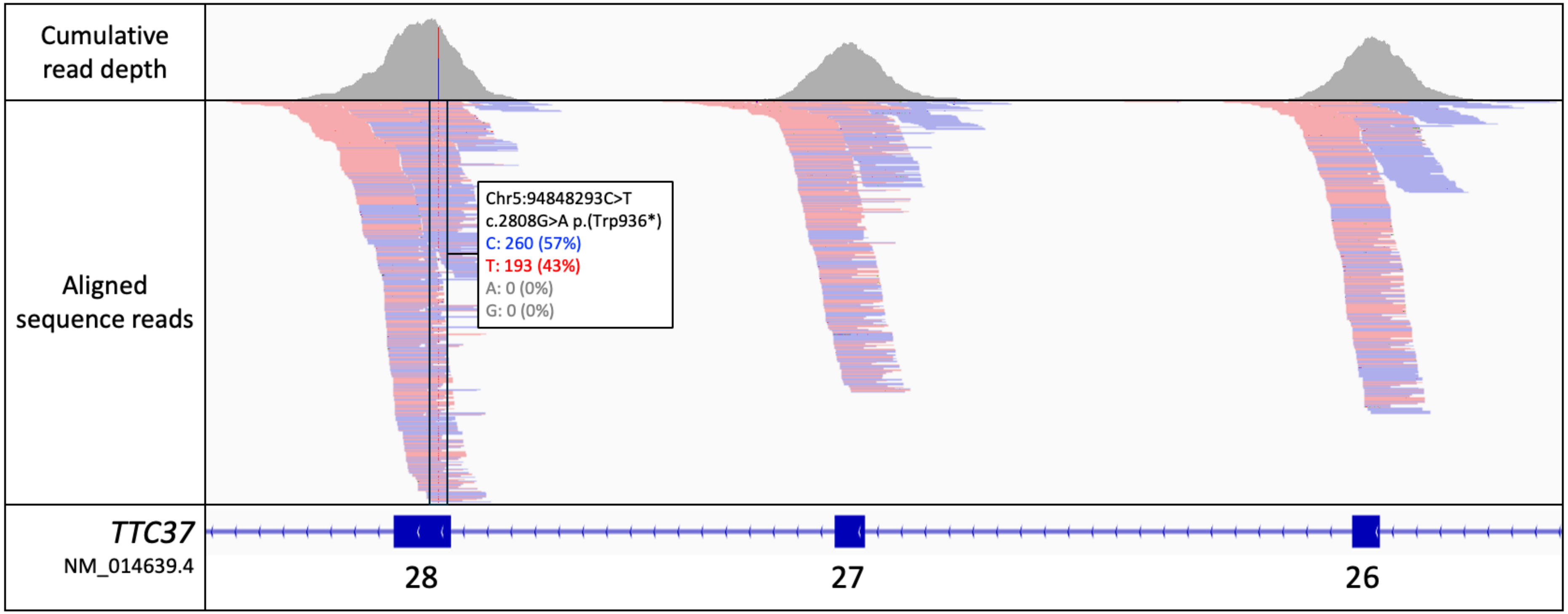

Figure 5.1: Short-read sequencing alignments showing the location of the single nucleotide variant c.2808G>A p.(Trp936*) in *TTC37*. The cumulative read depth at the variant site was 453; the variant allele fraction (C: 0.57, T: 0.43) indicates the variant is heterozygous. The black box in the aligned sequence reads track highlights the location of reads containing non-reference "T" nucleotides. Genomic coordinates are displayed for the human reference genome, build GRCh37 (in use at the time of original diagnostic testing). Aligned sequence reads are coloured based on their mapped read strand; Pink: "+" (plus) strand alignments, Blue: "-" (minus) strand alignments.

5.3.2 Long-read whole genome sequencing

Tetratricopeptide repeat domain 37, or *TTC37*, is a gene consisting of 42 exons, spanning 92 kb of chromosome 5 (Hartley *et al.,* 2010a). The gene is associated with tricho-hepato-enteric syndrome (THES), an autosomal recessive disorder characterised by intractable diarrhoea from infancy, trichorrhexis nodosa (weakness of the hair shaft, leading to hair thickening and breakages), and IUGR (Lee *et al.,* 2024). As these clinical features were consistent with the patient phenotype, it seemed highly likely that a second heterozygous variant may be present in this gene, warranting further analysis beyond the coding regions already examined. Due to the extensive size of the *TTC37* locus (preventing targeted analysis by long-range PCR enrichment), a long-read whole genome sequencing approach was used to assess the region for a potential second variant.

Prior to preparing a long-read sequencing library the fragment size distribution of the genomic DNA sample was first assessed using a Femto Pulse. A fragment size distribution of approximately 1-250 kb was observed (Figure 5.2A). To achieve an optimal data yield and minimise blocking of the nanopores by ultra-high molecular weight fragments, the genomic DNA was first sheared using a G-tube, which utilises centrifugal force to draw HMW DNA through a narrow orifice and fragments to a range of 6-20 kb. A short-read elimination step was then performed to aid the removal of DNA fragments shorter than 10 kb. The specimen was again reviewed using a Femto Pulse to observe the change in fragment profile; this resulted in a sample that had an average fragment distribution of 10-50 kb, with a peak at 16 kb (Figure 5.2B). 1 µg of the size selected material was used for long-read library preparation. The yield of this initial library preparation was 5.12 ng/µl.

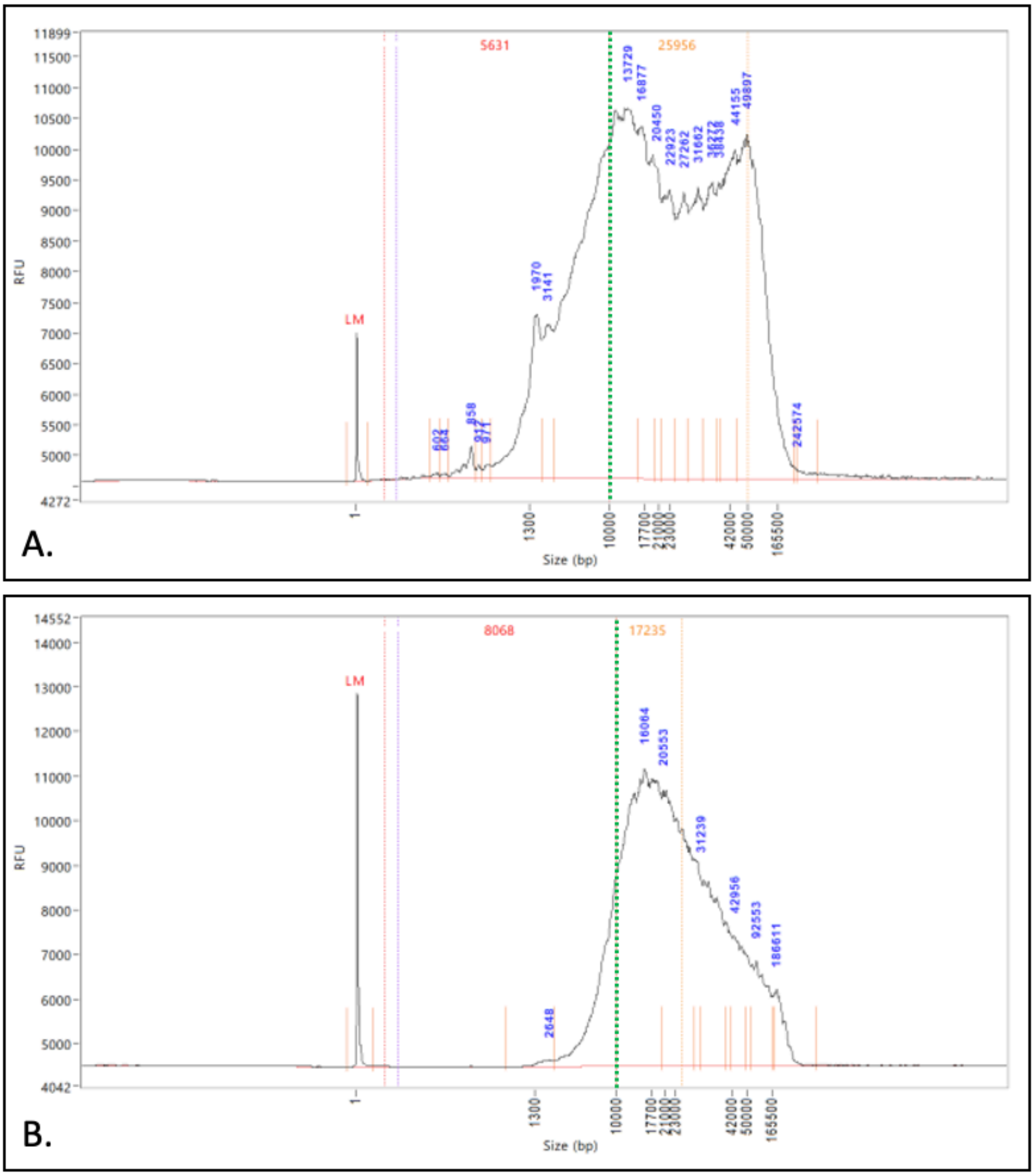

Figure 5.2: Femto Pulse traces generated to assess DNA quality for long-read whole genome sequencing. (A) The fragment size distribution of the genomic DNA sample of the TTC37 proband (prior to shearing or SRE treatment). **(B)** The processed genomic DNA sample following G-tube shearing and treatment with SRE. Note the change in the fragment profiles; the proportion of high molecular weight molecules was reduced (the effect of shearing), and the majority of fragments less than 10 kb (defined by the dashed vertical green line) were removed (the effect of SRE treatment). The *x*-axis is log scaled and the upper marker is 165,500 bp.

When reviewing sequencing yield over time, the performance and availability of the nanopores decrease; this is largely due to the blocking of the pores by residual high molecular weight fragments. The effect can be visualised as a decrease in sequencing activity, shown by the green bars of the histogram in Figure 5.3. In order to boost run performance (by increasing the data yield), it is possible to nuclease flush the flow cell. During this process the blocked fragments are removed and fresh library is re-loaded onto the flow cell. As more than half of the library was loaded (10 fmol, 98.55 ng) during the initial run setup, a second library was prepared using a further 1 µg of size selected DNA to enable the flow cell to be flushed and reloaded. As the second library's concentration was 13.2 ng/µl it was possible to use it to flush and reload the flow cell twice (this occurred at approximately 24 and 48 hours of sequencing), with the increase in sequencing activity shown in Figure 5.3. The total data yield from the flow cell was 104.44 Gb and the N50 value of the sequence reads was 15.47 kb. The size distribution of the sequencing reads is shown in Figure 5.4; despite a bias towards sequencing shorter reads, the profile was comparable to that seen on the Femto Pulse analysis of the input material (Figure 5.2B).

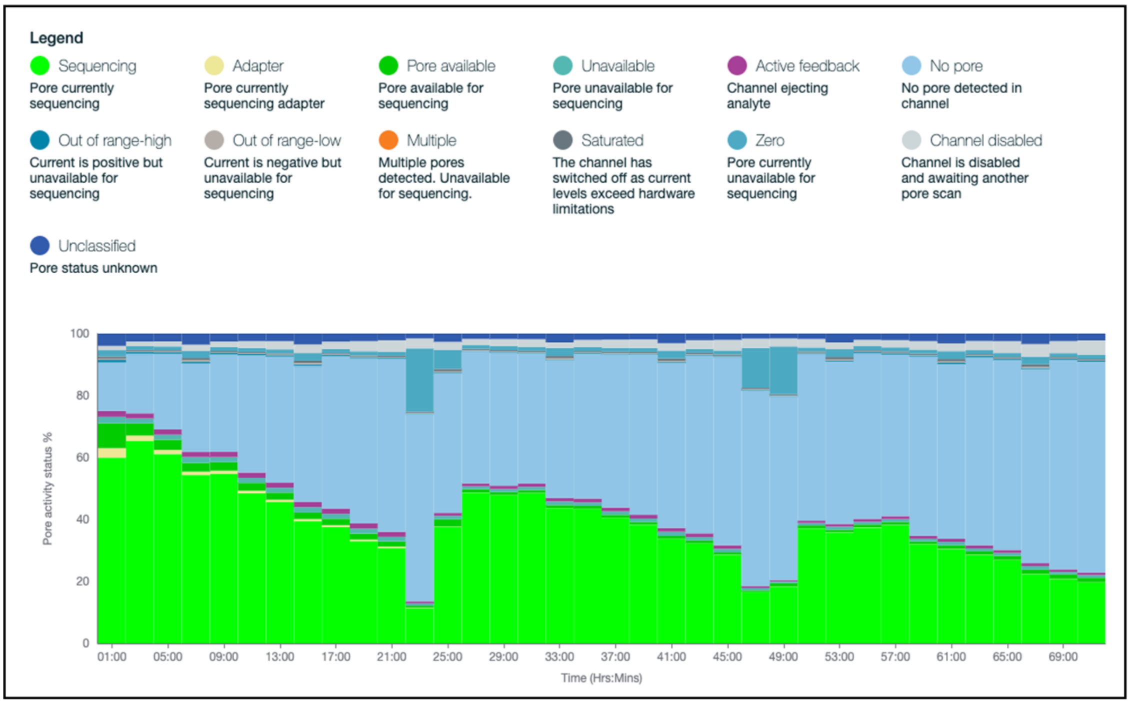

Figure 5.3: Flow cell activity during the PromethION sequencing run. The increase in sequencing activity (green bars) at 25 hours and 50 hours is due to a nuclease flush and sample reload having been performed.

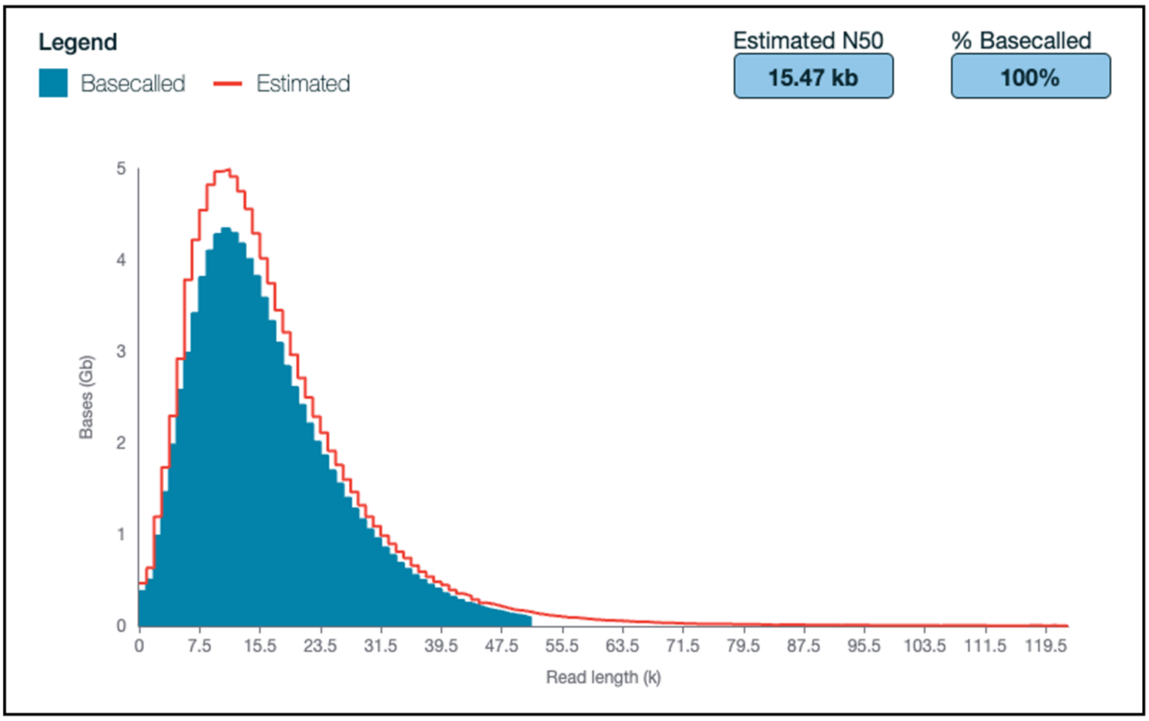

Figure 5.4: Read length distribution following long-read sequencing using a PromethION flowcell.

Whole-gene analysis of *TTC37* using long read whole genome sequencing data identified three homozygous and three heterozygous variants (Table 5.2). The three heterozygous variants clustered within a range of 5.53 kb around the mid-point of the gene (Figure 5.5), with one of these being the previously detected chr5:95512589 C>T (c.2808G>A p.(Trp936*). The reads were grouped into reference and non-reference nucleotides at this position in order to phase the identified heterozygous variants into their respective haplotypes. Intronic variant chr5:95515674T>C, c.2634+679A>G was observed in *trans* with the previously detected pathogenic mutation, making it the only candidate variant that could account for a missing "second hit" mutation (Figure 5.6). Automated analysis using the Clair3 variant caller did not identify any further candidate variants in *TTC37* and was consistent with the manual inspection of the sequence reads using the IGV. Furthermore, analysis of this data using the structural variant caller Sniffles2 did not identify any additional variants. Average read depth across the *TTC37* gene was approximately 35×.

Table 5.2: Variants identified in *TTC37* following long-read whole genome sequencing.

g.Nomen	c.Nomen	p.Nomen	Zygosity	Location	dbSNP identifier	gnomAD* allele frequency
g.95466983A>G	c.4620+883T>C	p.?	Hom	Intron 42	rs2731817	99.955
g.95493954A>G	c.3802+728T>C	p.?	Hom	Intron 36	rs1746686172	-
g.95510148G>C	c.2921-461C>G	p.?	Het	Intron 28	rs537638358	0.003
g.95510212A>G	c.2921-525T>C	p.?	Hom	Intron 28	rs62365053	5.203
g.95512589C>T	c.2808G>A	p.(Trp936*)	Het	Exon 28	rs534237033	0.006
g.95515674T>C	c.2634+679A>G	p.?	Het	Intron 25	-	-

Genomic coordinates are provided for chromosome 5 according to the human reference genome build GRCh38. Coding nomenclature is provided according to transcript NM_014639.4. g.Nomen: Genomic nomenclature. c.Nomen: Coding nomenclature. p.Nomen: Protein nomenclature. Hom: Homozygous. Het: Heterozygous. *: gnomAD version v.2.1.1. dbSNP database: https://www.ncbi.nlm.nih.gov/snp/

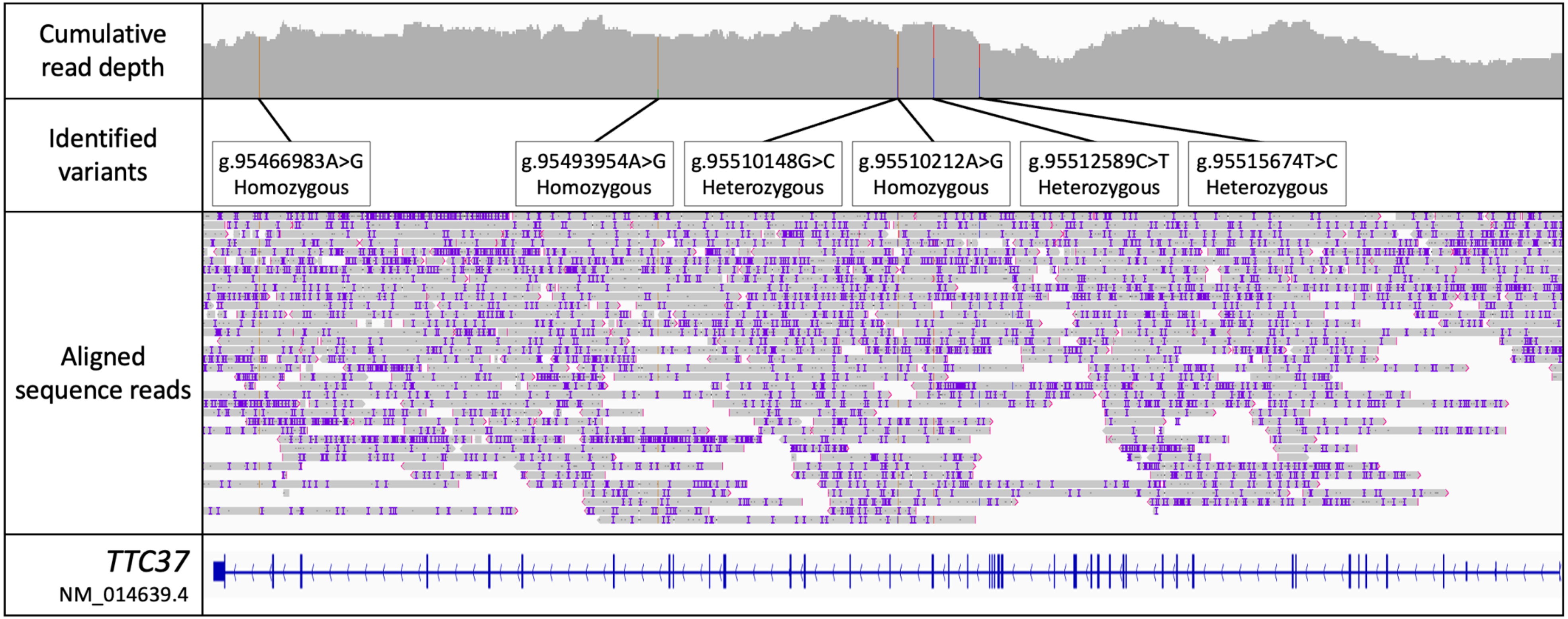

Figure 5.5: Unphased long-read sequence reads aligned at the *TTC37* locus. Three homozygous and three heterozygous single nucleotide variants were observed (complete variant nomenclature is recorded in Table 5.2). The *y*-axis of the cumulative read depth chart is scaled 0-40×. The *TTC37* gene structure is defined in blue; horizontal blue lines represent introns, vertical blue lines represent exons, arrows indicate the direction of transcription (*TTC37* is encoded on the antisense strand).

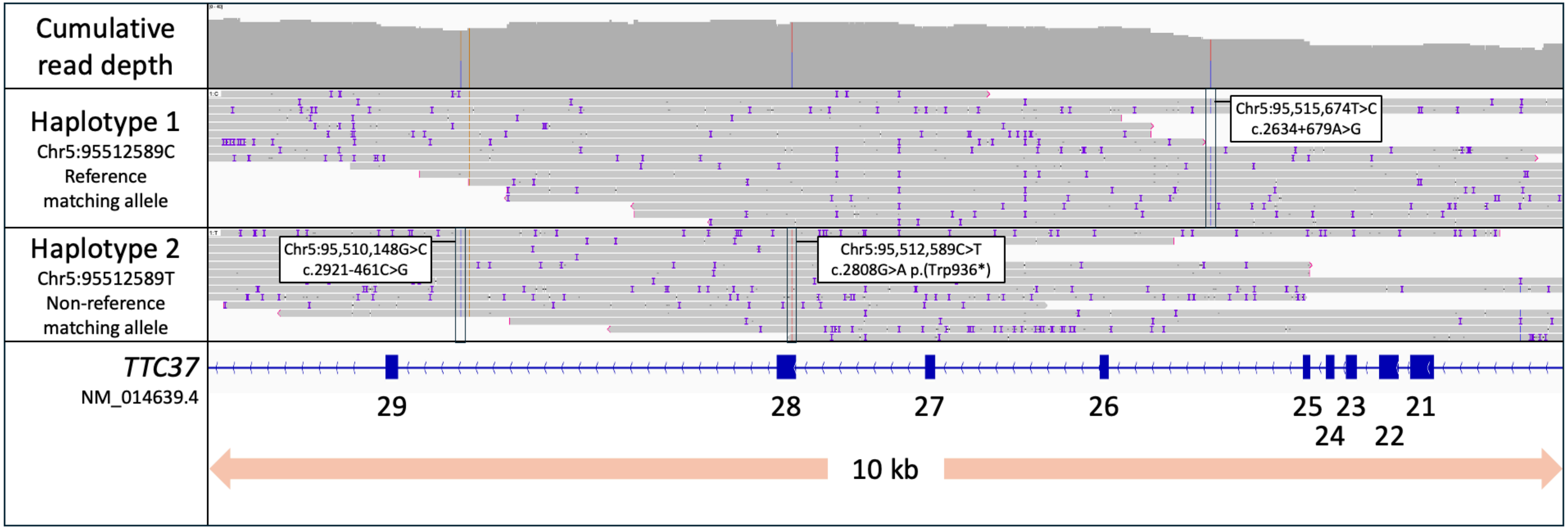

Figure 5.6: Phased haplotypes for the three variants identified in *TTC37*; haplotypes were established using the bases at position Chr5:95512589. Of two additional variants that were identified by this analysis only one, Chr5:95515674T>C (c.2634+679A>G) was identified to be on a different parental haplotype to the originally identified heterozygous pathogenic variant Chr5:95512589C>T (c.2808G>A; p.(Trp936*)). Genomic coordinates are provided for the human reference genome build GRCh38. The *y*-axis cumulative read depth is scaled 0-40×. *TTC37* is encoded on the antisense strand. Exons are depicted as blue squares and numbered according to transcript NM_014639.4.

5.3.3 *In silico* splicing prediction for c.2634+679A>G

To assess the functional effect of the c.2634+679A>G variant *in silico* splicing analysis was performed using SpliceAI; the tool generates a delta score, ranging from 0 to 1, on the probability of a nucleotide change affecting splicing within a given window (by default this is +/- 500 bp). For the *TTC37* variant c.2634+679A>G a probability score of 0.98 was returned for a donor site gain located 5 bp downstream. This suggested a high likelihood that the variant affected splicing. To further verify the splicing prediction the variant was inputted into Alamut Visual (v1.3) which hosts an additional 4 splicing prediction tools. Three of these tools (SpliceSiteFinder-like, MaxEntScan and NNSPLICE) strongly predicted the same donor splice gain that was predicted by SpliceAI (Figure 5.5). While the *in-silico* evidence was strongly suggestive of a functional effect, (with the c.2634+679A>G variant therefore accounting for the missing second-hit pathogenic variant in this patient), the clinical ACGS guidelines for variant classification do not permit interpretation of non-canonical splicing variants as pathogenic based on computational evidence alone (Durkie *et al.,* 2024). To experimentally determine the functional effect of the c.2634+679A>G variant RNA analysis was performed.

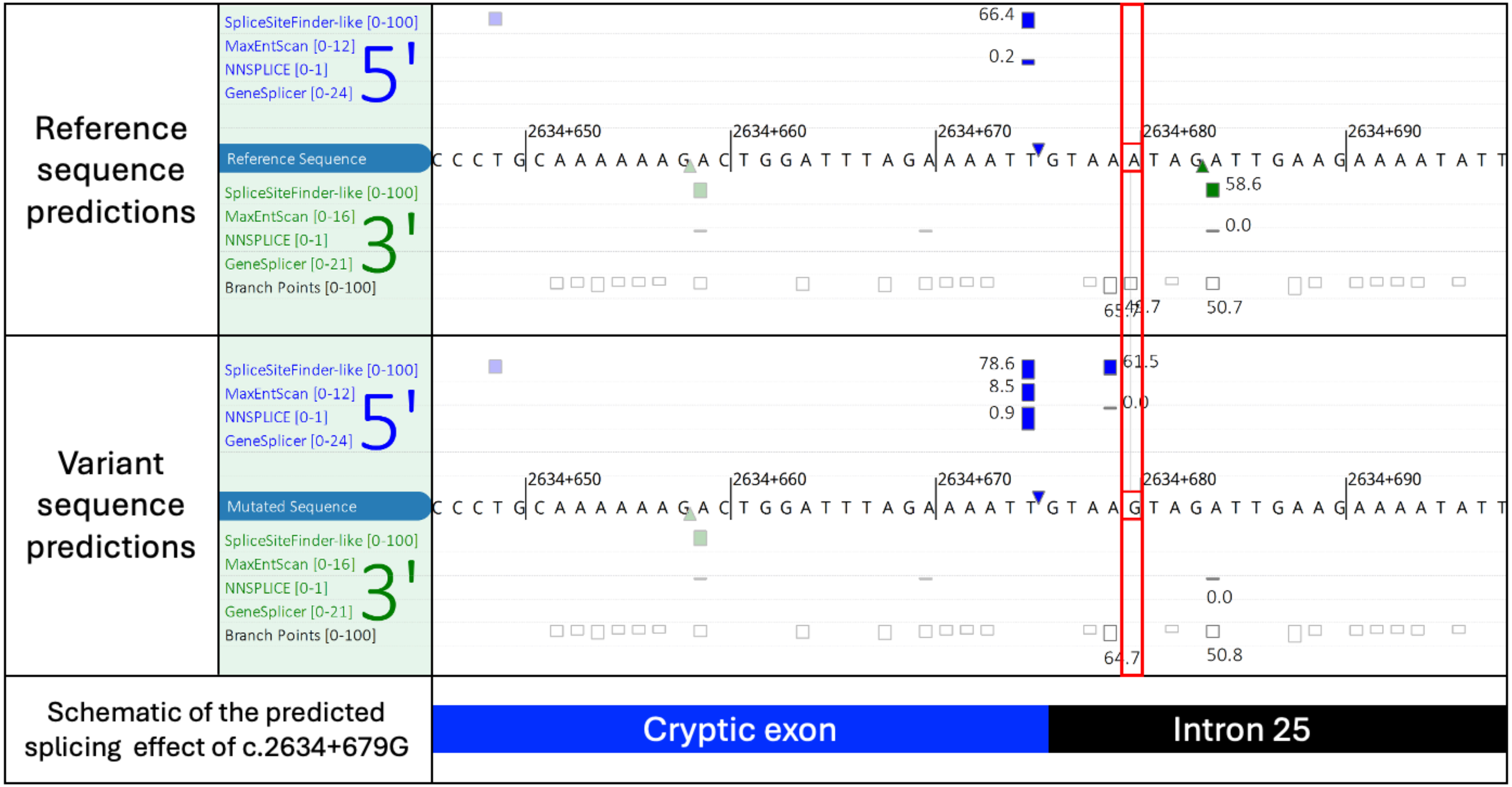

Figure 5.7: ***In silico*** **splicing predictions for the intronic variant c.2634+679A>G.** Reference and variant nucleotides are highlighted within the red box in the upper and lower tracks respectively. Results from four splicing prediction tools are reported (SpliceSiteFinder-like, MaxEntScan, NNSPLICE and GeneSplicer); each tool's numerical reporting range is displayed in brackets. A prediction is provided for the creation of both 5' and 3' splice sites.

5.3.4 RNA Extraction

RNA was extracted from peripheral blood of the proband using a PAXgene RNA Blood Kit. The RNA was quantified using the Qubit RNA BR Assay, with a measured concentration of 56.2 ng/µl. The RNA was also assessed using an RNA ScreenTape on the TapeStation (Figure 5.8). This analysis generated an RNA integrity number equivalent (RIN^e), which is a measure of the quality of the RNA sample (1 represents degraded RNA whereas 10 represents intact, high-quality, RNA). The RNA ScreenTape measurement generated a RIN^e of 7.9 for this sample.

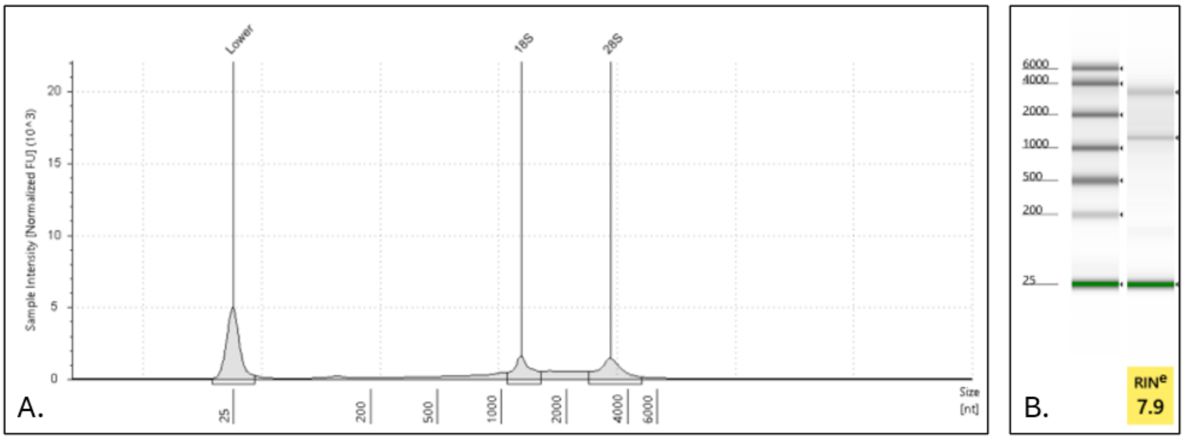

Figure 5.8: RNA extraction quality control assessment using TapeStation RNA ScreenTape. (A) The electropherogram displays the 18S and 28S fragment peaks. **(B)** The overall fragment distribution across the region is measured against a ladder (of known fragment sizes) to calculate the RIN^e value.

5.3.5 Short-read whole transcriptome RNA sequencing

To verify the functional effect of the predicted novel splice site, whole transcriptome RNAseq was performed. Short-read sequencing summary metrics, generated from one lane of a NovaSeq SP flow cell are detailed in Table 5.3; these are consistent with the instrument specification for a 300-cycle cartridge. The aligned sequence reads mapping to *TTC37,* and 10 kb of flanking sequence were manually compared for the patient and a healthy control using the IGV, with reference to the RefSeq annotation of the *TTC37* transcript. This revealed a novel cryptic exon between exons 25 and 26 (Figure 5.9). The cumulative coverage plot for the novel exon had a total depth that was approximately half the height of the adjacent exons, suggesting (as expected) only one allele was affected. A Sashimi plot of these data identified the number of novel exon-exon junction spanning reads to be approximately half those of the reference exons (Figure 5.10). The impact of the novel cryptic exon on the reading frame was next assessed. This revealed that the novel exon is predicted to introduce a premature termination codon (Figure 5.11).

Table 5.3: Sequencer run metrics for short-read RNAseq library.

Read length	Yield (Gb)	Proportion of bases >=Q30 (%)	Total read pairs
151	138.9	92.1	484,740,023

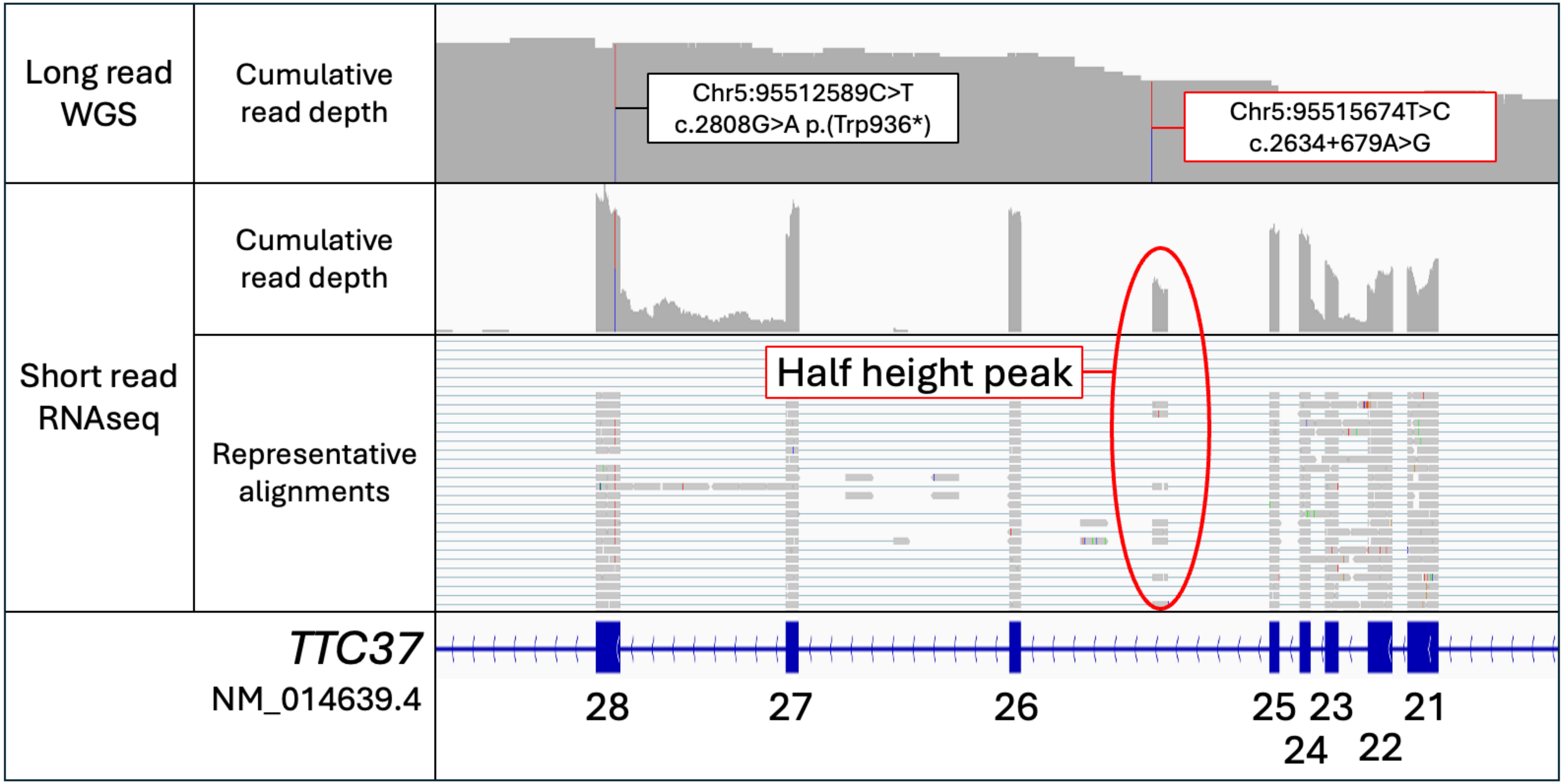

Figure 5.9: Short-read whole transcriptome RNAseq supports a predicted cryptic exon. The cryptic exon is circled red, note that the relative cumulative read depth is approximately half the height of that at the adjacent exons suggesting that (as expected), only one allele contains the novel exon. *TTC37* is encoded on the antisense strand; arrows show the direction of transcription. Blue squares denote exons and are numbered according to transcript NM_014639.4. Genomic coordinates are reported for the human reference genome, build GRCh38.

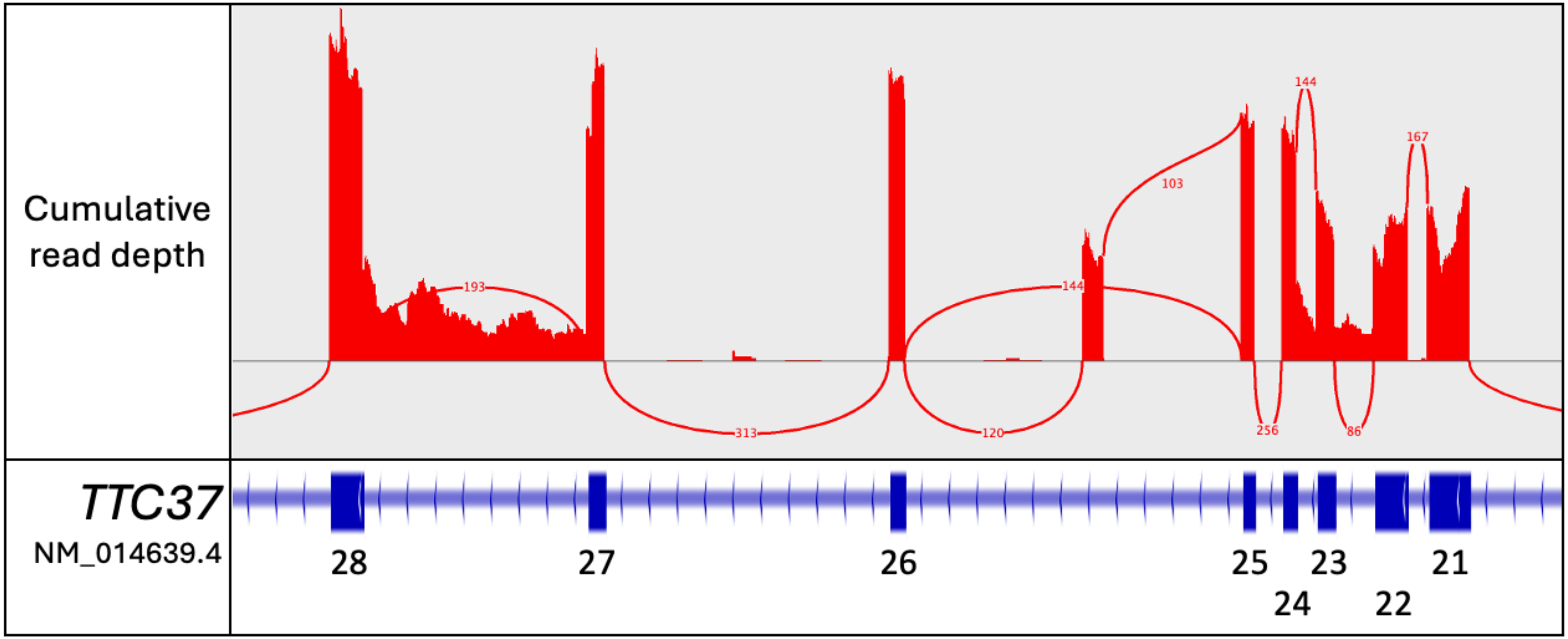

Figure 5.10: Sashimi plot showing the number of *TTC37* exon-exon junction spanning reads. The count of junction spanning reads is displayed above the corresponding exon. Approximately half the number of reads map between exons 25 and 26. Note the half-height peak in intron 25 which corresponds to the novel cryptic exon sequence. The cumulative coverage *y*-axis is scaled 0-400x.

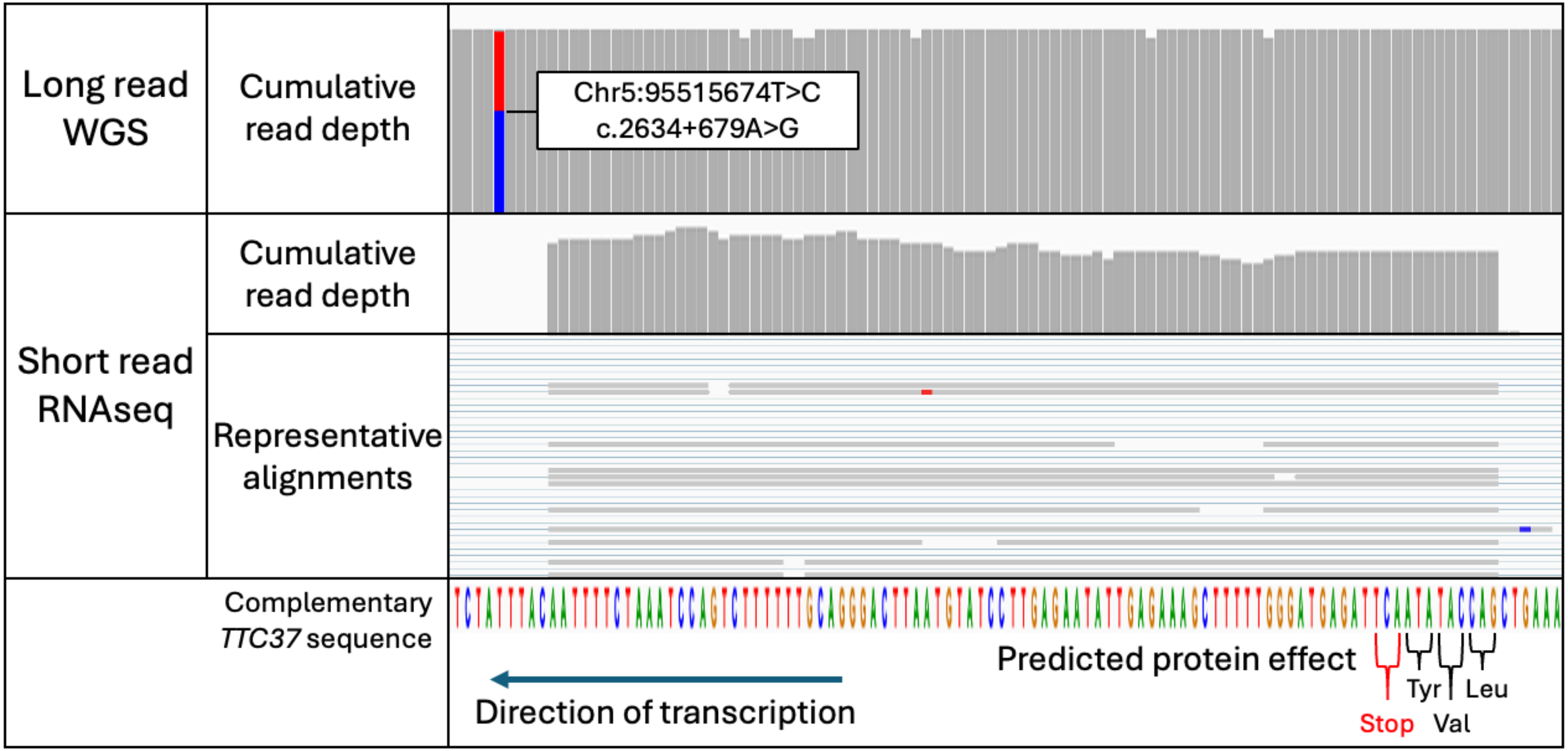

Figure 5.11: Detailed analysis of the short-read alignments reveals that the cryptic exon boundaries are consistent with the *in-silico* prediction. See Figure 5.5 for the *in-silico* prediction (note that the image is displayed left-to-right from a transcript perspective, compared to the sense strand genome alignment in the current figure). The novel-exon is predicted to create a stop codon leading premature termination of the translated protein.

5.3.6 Long-read whole transcriptome cDNA sequencing

While the short read RNAseq dataset provided functional evidence for the effect of the c.2634+679A>G variant, it was not possible to confirm that the novel cryptic-exon containing transcript was in trans with the c.2808G>A; p.(Trp936*) variant. We therefore sought to use long-read RNAseq to determine whether phase could be observed at the transcript level. A globin-depleted cDNA long-read RNAseq library was prepared using kit cDNA-PCR Sequencing Kit v14. The cDNA profile was first assessed using a TapeStation D5000 ScreenTape (Figure 5.12). The yield of the prepared library was 4.8 ng/µl (as measured by the Qubit using a DNA BR Assay) and this was sequenced on a single PromethION R10.4.1 flow cell. The 72-hour run yielded 22.69 Gb of sequence data from 25.82 million reads with an N50 of 890 bp. Alignment of these reads to the human reference genome (build GRCh38) using the long-read splice aware aligning parameters of minimap2 identified a single informative read which spanned from exon 22 to 43 (Figure 5.13). The sequence read was from the cryptic-exon containing allele and the sequence contained a reference-matching nucleotide at position c.2808 (chr5:95512589). This supported the two pathogenic variants being in *trans* at the level of transcription.

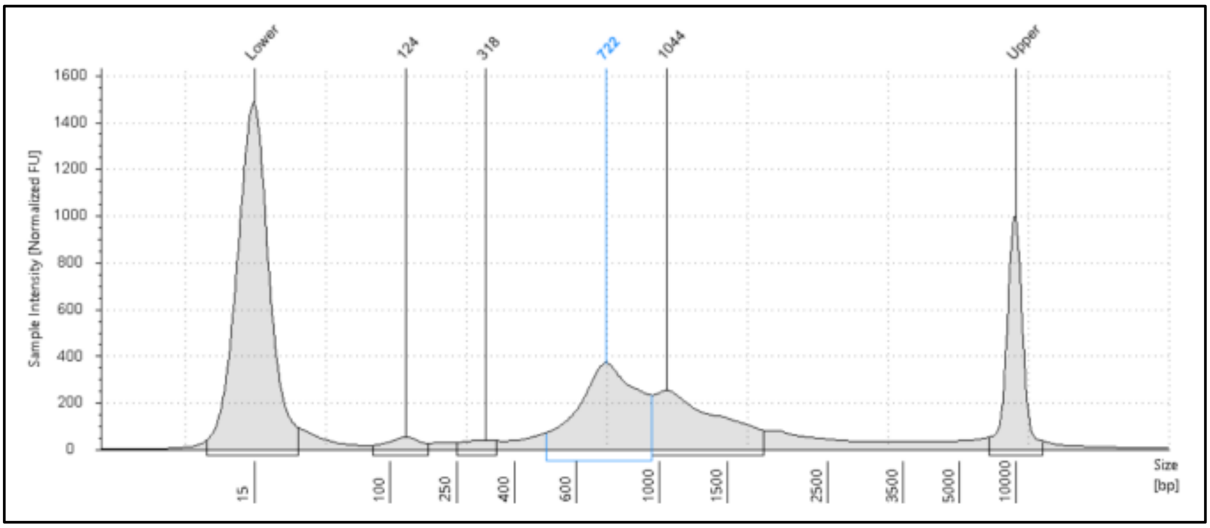

Figure 5.12: cDNA quality control assessment using a TapeStation D5000 ScreenTape. The majority of cDNA fragments are between 500-1,500 bp; note the peak at 722 bp.

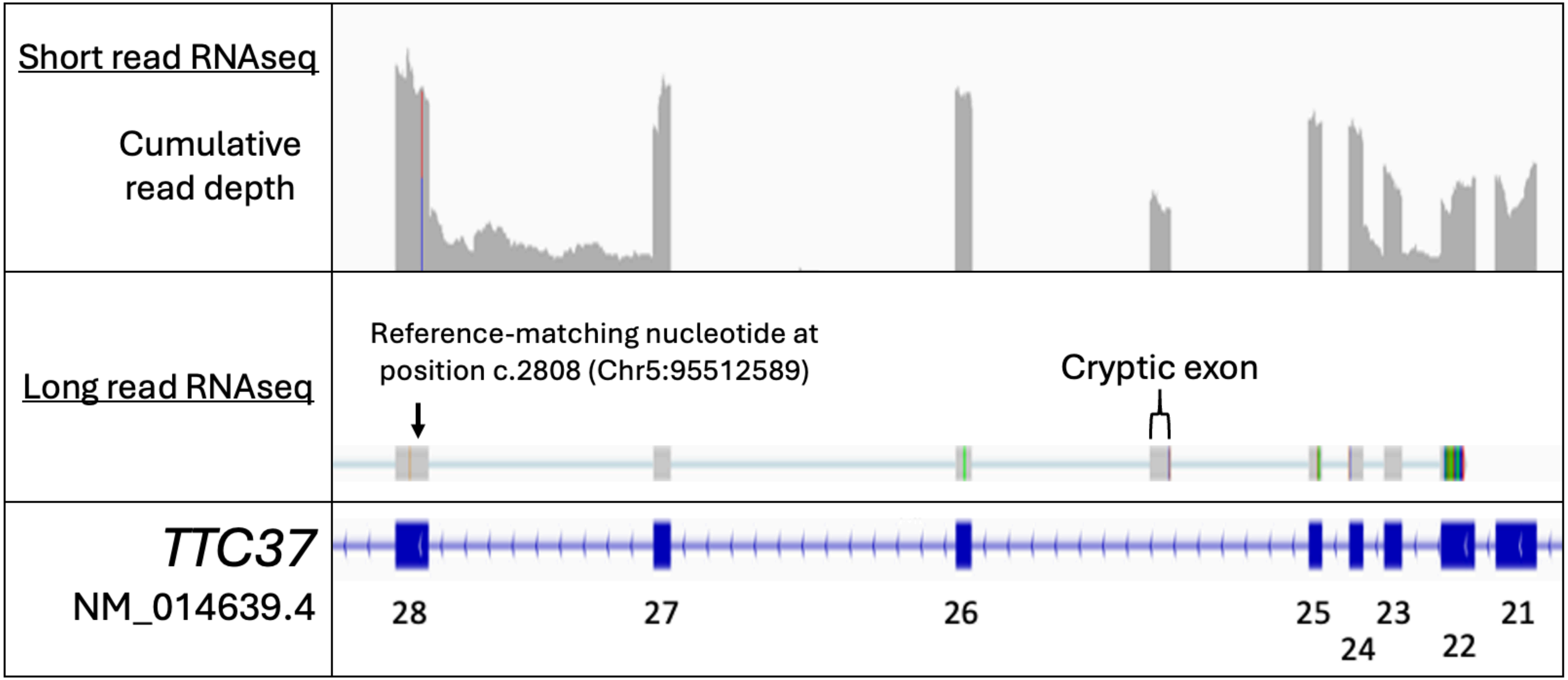

Figure 5.13: A single long read RNAseq alignment was identified which demonstrated that the cryptic exon sequence is cis with a reference matching nucleotide at transcript position c.2808.

5.3.7 Assay for segregation analysis and cascade screening

To verify the identified variants by Sanger sequencing and establish an assay for segregation testing among additional family members, a PCR amplicon was optimised. These data confirmed that the c.2808G>A p.(Trp936*) variant was paternally inherited (Figure 5.14A) and the c.2634+679A>G variant was maternally inherited (Figure 5.14B).

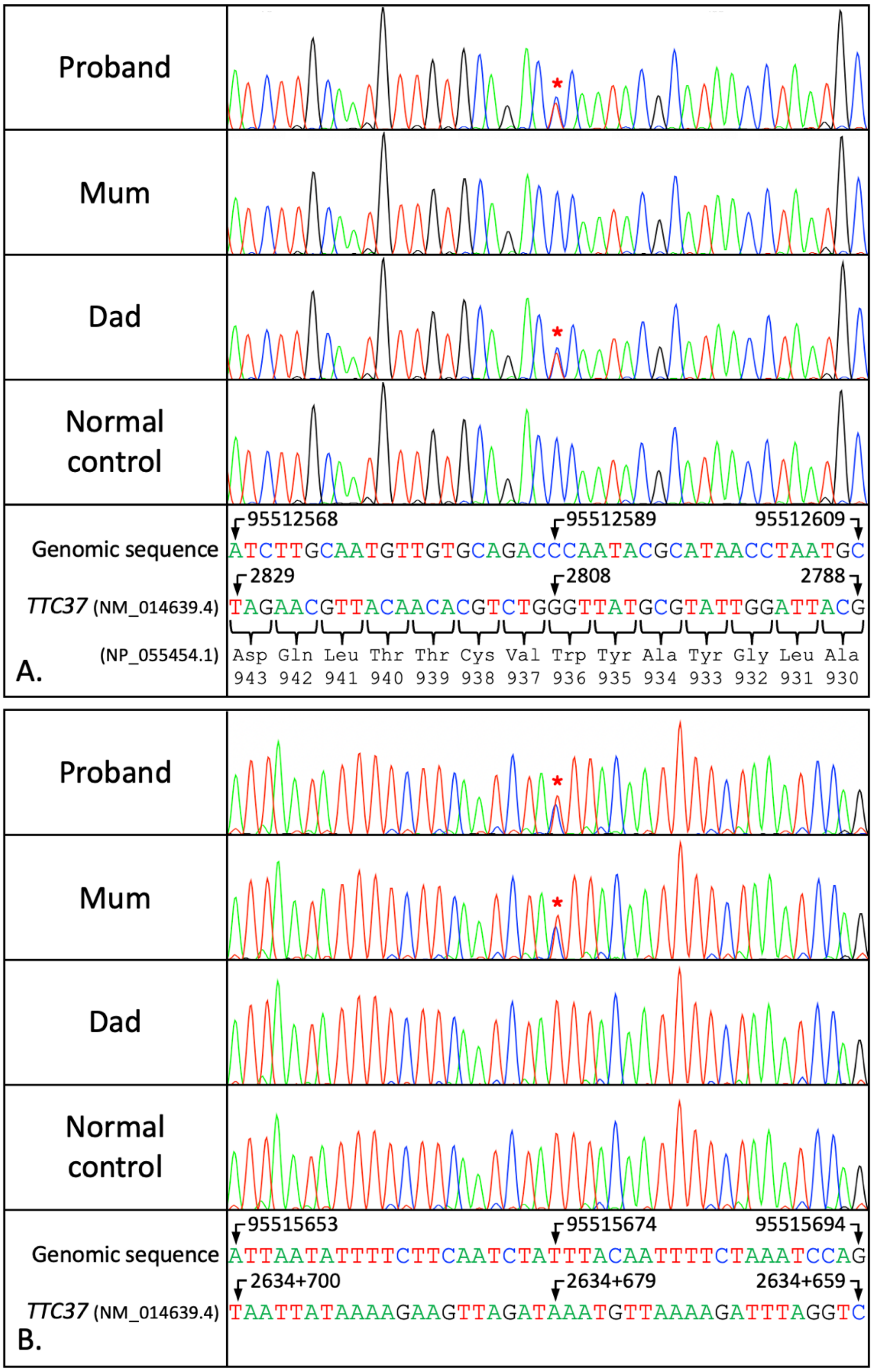

Figure 5.14: Segregation analysis of the pathogenic *TTC37* variants. **(A)** The c.2808G>A p.(Trp936*) (chr5:g.95512589C>T) heterozygous variant is confirmed in the proband and determined to be inherited from the patient's father. **(B)** The c.2634+679A>G (chr5:g.95515674T>C) heterozygous variant is confirmed in the proband and determined to be inherited from the patient's mother. Genomic coordinates are reported for human reference genome build GRCh38. Transcript number is according to NM_014639.4. Variant positions are highlighted with the red star.

5.4 Case 2: Results

5.4.1 Standard-of-care investigation

A couple were referred for genetic testing following the identification of fetal abnormalities by ultrasound. These included encephalocele (a neural tube defect where brain tissue protrudes outside of skull) and cystic kidneys. Rapid prenatal exome sequencing was requested; this was performed using a fetal chorionic villus sample (CVS), in combination with parental DNA samples. A targeted analysis was performed that included 1,239 genes from the "R21" Fetal Anomalies gene panel (v3.0) (https://nhsgms-panelapp.genomicsengland.co.uk/panels/478/v3.0). The heterozygous variant, c.1046T>C p.(Leu349Ser), which had previously been reported to be pathogenic in the medical literature, was identified in exon 10 of the *TMEM67* gene. Assessment of parental DNA samples (sequenced at the same time as the CVS) demonstrated that the variant was maternally inherited (Figure 5.15).

Evidence supporting the pathogenicity of the c.1046T>C variant includes its presence in only heterozygous form, and at very low levels (66 alleles out of 1606212 alleles), in the population allele frequency database gnomAD (https://gnomad.broadinstitute.org/). By contrast, the c.1046T>C variant has been reported in homozygous form (Iannicelli *et al.,* 2010) and in combination with pathogenic/likely pathogenic variants in multiple individuals with TMEM67-associated disorders (Khaddour *et al.,* 2007). Bioinformatic analysis predicts a strong pathogenic effect for the amino acid substitution of a hydrophobic leucine residue for a polar uncharged serine; this is reflected in a REVEL score of 0.814 and a CADD score of 28.5. *In vitro* functional studies of the extracellular N-terminal region of TMEM67 have demonstrated that the p.Leu349Ser variant abolished binding to Wnt5a (Abdelhamed *et al.,* 2015).

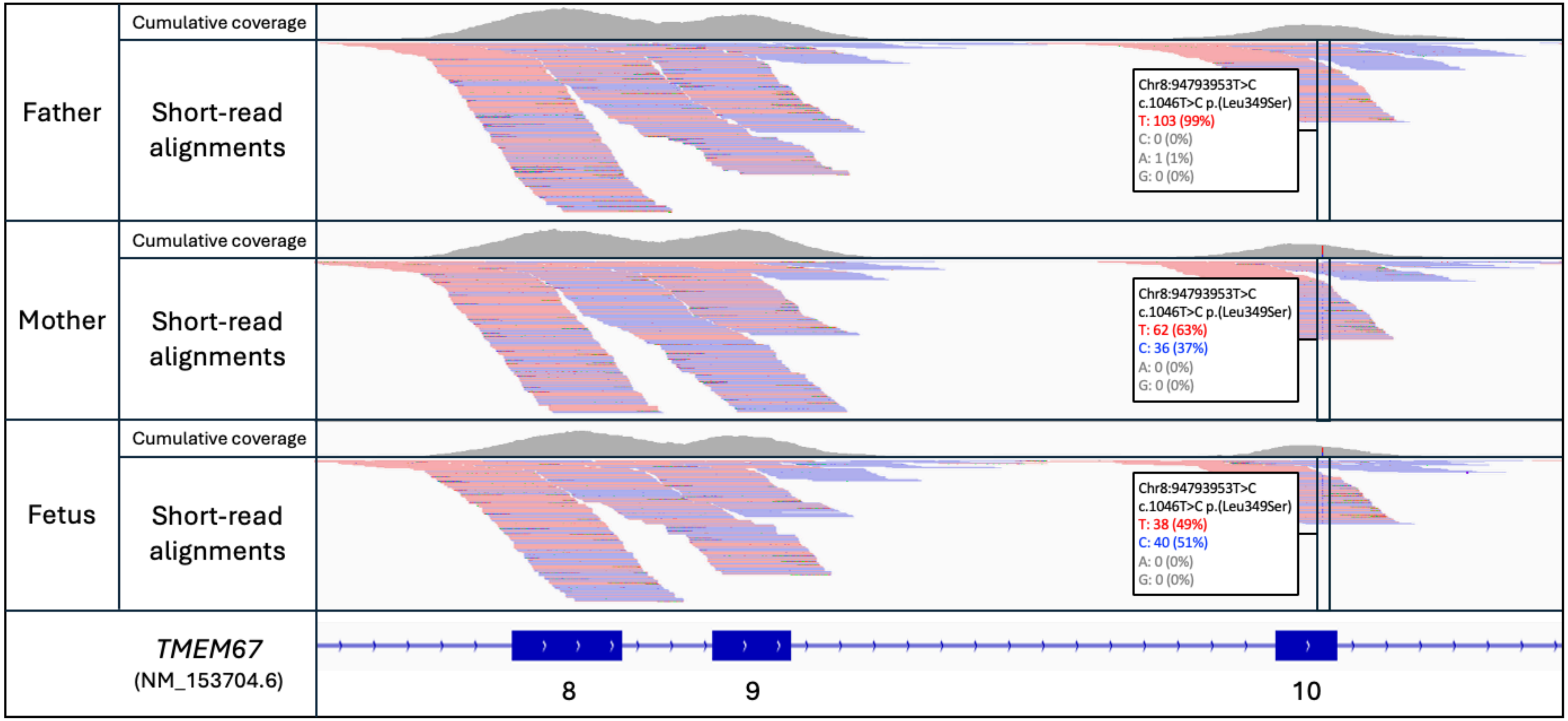

Figure 5.15: Identification of the heterozygous pathogenic *TMEM67* variant c.1046T>C p.(Leu349Ser) in the fetus by exome sequencing. Short-read sequences were aligned to build GRCh37 of the human reference genome (in use in the diagnostic pathway at the time of testing. cDNA numbering is according to transcript NM_153704.6; exons are displayed as blue rectangles and numbered accordingly. The chorionic villus sample from the fetus is heterozygous for the c.1046T>C variant which was inherited from the mother. The *y*-axis of the cumulative coverage plot is scaled 0-250x. Reads are coloured according to the strand on which they were mapped.

5.4.2 Long-read whole genome sequencing

Pathogenic mutations in Transmembrane Protein 67 (*TMEM67*) are associated with Meckel-Gruber Syndrome (OMIM: 607361), an autosomal recessive disorder characterized by renal cystic dysplasia and developmental defects of the central nervous system, in particular encephalocele (Smith *et al.,* 2006). Given that this phenotype was consistent with the phenotype that was detected by ultrasound, it was plausible that a second pathogenic variant, which was required to confirm a diagnosis of Meckel-Gruber Syndrome, may be present in a region of the *TMEM67* gene that had not been screened by the routine standard-of-care analysis. We sought to test this hypothesis by long-read whole genome sequencing.

Long-read whole genome sequencing requires a large mass of DNA. However, due to the fetal DNA having been extracted from a chorionic villus sample (CVS), there was limited material available. For this reason, and because the previously characterised *bona fide* pathogenic variant had been detected in the mother, the father's sample was prepared for long-read sequencing.

Prior to generating a long-read sequencing library the fragment size distribution of the genomic DNA sample obtained from the father was assessed using a Femto Pulse. A fragment size distribution of approximately 1-200 kb was observed (Figure 5.16). As there were minimal fragments detected below 10 kb, and the majority of the sample was within the desired range of 10-50 kb, the sample was processed for sequencing without shearing or SRE being performed. 1 µg of paternal DNA was used to create a long read ONT library which had a final concentration of 24.2 ng/µl.

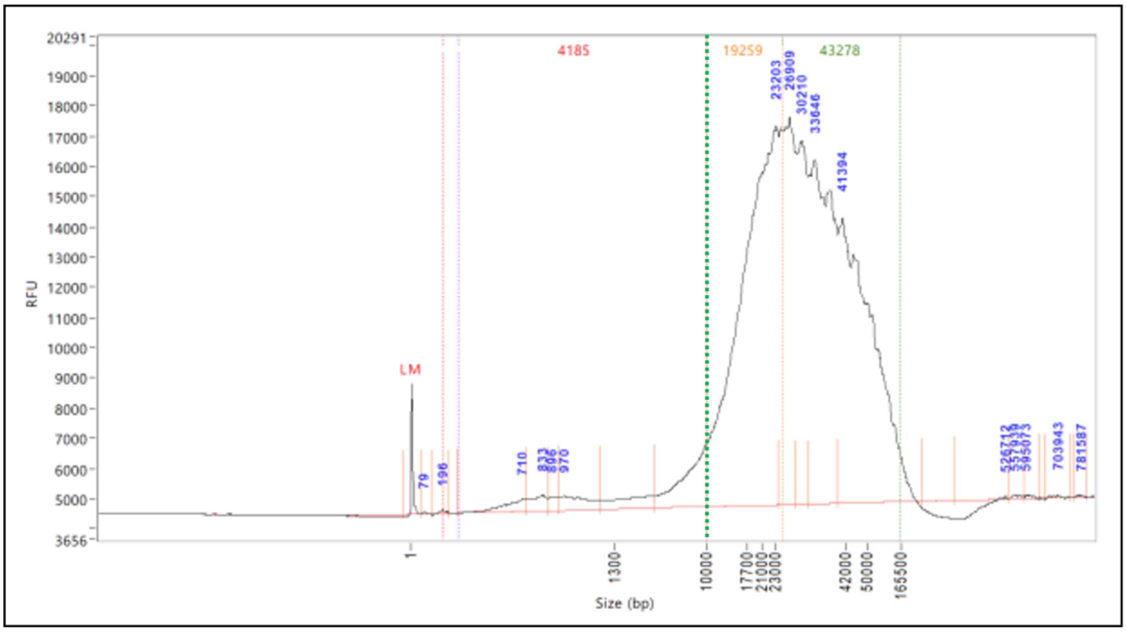

Figure 5.16: Femto Pulse trace generated as part of the assessment of paternal DNA quality for long-read whole genome sequencing. The fragment size distribution of the genomic DNA sample. The major peak was determined to be approximately 25 kb, with very few DNA fragments detected below 10 kb (defined by the dashed vertical green line). The sample profile was interpreted to be suitable for long-read sequencing, without shearing or treatment using short read eliminator reagents.

The initial loading of the PromethION flowcell used 6.4 µl of library, based on the average size being estimated at 25 kb from the Femto Pulse trace (154 ng therefore equated to 10 fmol of DNA). As with Case 1, nuclease flushes and sample reloading was performed after approximately 24 and 48 hours of sequencing. The purpose of this was to clear pore blocked pores and boost sequencing yield. As the PromethION provides real-time feedback of fragment read lengths during sequencing, it was apparent that the read-length N50 was approximately 11 kb after 24 hours of sequencing (the overall N50 is a cumulative value that fluctuates over the course of the run). The quantity of library required to achieve 10 fmol of loading, for each reload, was therefore adjusted based on this value; consequently 2.8 µl (67.8 ng) was loaded at 24 and 48 hours, with a corresponding increase (boost) in sequencing activity visible in Figure 5.17.

The read length distribution of sequenced fragments is displayed in Figure 5.18; the majority of reads are <20 kb. As real-time basecalling was not performed for this run, the read length distribution for this sample is estimated. Nevertheless, when comparing the read length distribution of Case 1 (Figure 5.4), which was basecalled in real-time, it was observed that the actual read length distribution does closely mirror the estimated distribution; this suggests that the estimated skew towards shorter reads is accurate. This interpretation is also reflected in the final N50 value for the run, which was 11.85 kb. The total sequence yield that was achieved for the run was 134.06 Gb.

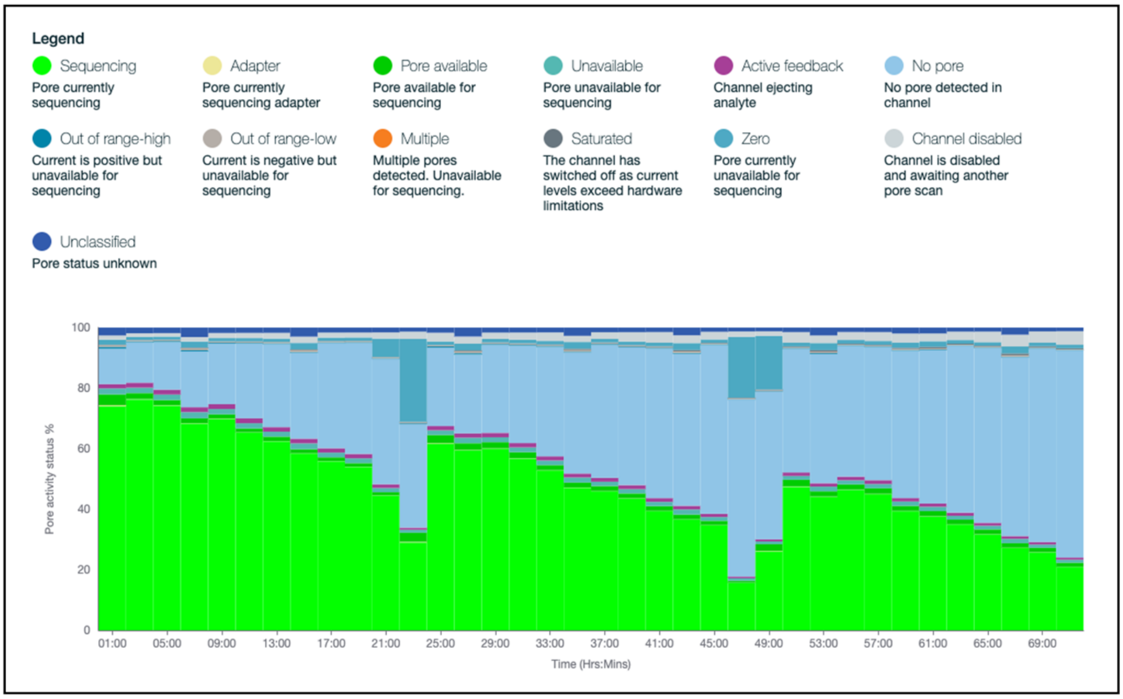

Figure 5.17: Flowcell activity during the PromethION sequencing run of paternal DNA from *TMEM67* case. The increase in sequencing activity (green bars) at 25 hours and 50 hours is due to a nuclease flush and sample reload having been performed.

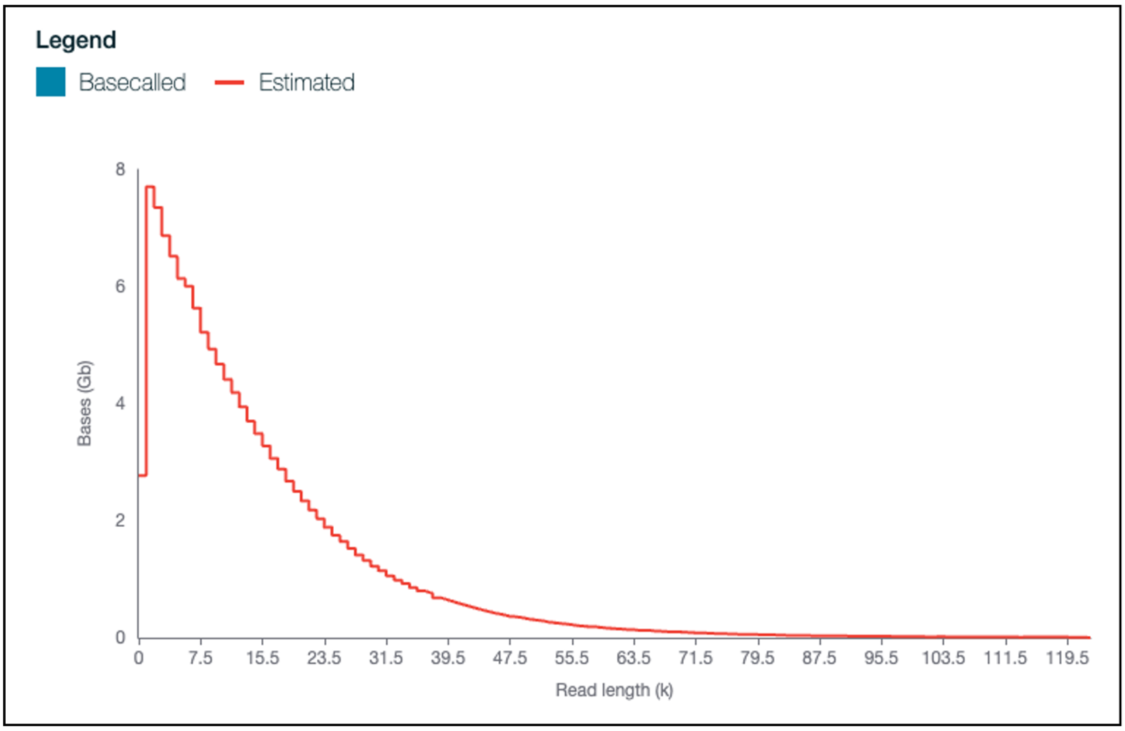

Figure 5.18: Read length distribution following sequencing of paternal DNA from *TMEM67* case. Note that unlike in the previous case, basecalling was not performed in real-time; the presented read length distribution is therefore estimated from the raw data.

Long-read sequences were mapped to the human reference genome; those mapping to the *TMEM67* locus were reviewed using the IGV. A cluster of soft-clipped reads in intron 25 (defined as reads whose entire sequence could not be continually aligned to a single locus in the reference genome), suggested the presence of a structural variant (Figure 5.19B). The soft-clipped reads were extracted from the BAM file and submitted to the BLAT webserver (Kent, 2002). This process allowed the genomic coordinates of the unknown “soft-clipped” sequences to be identified. The structural variant was resolved to be a 3.4 Mb inversion, whose opposite breakpoint intersected the *LINC00534* gene (Figure 5.20). Re-inspection of the upstream *LINC00534* breakpoint revealed a cluster of soft-clipped reads (Figure 5.19A). Notably, the breakpoint intersects an MLT2A2 long terminal repeat element, the low complexity of these genomic features can give rise to structural variation due to non-allelic homologous recombination.

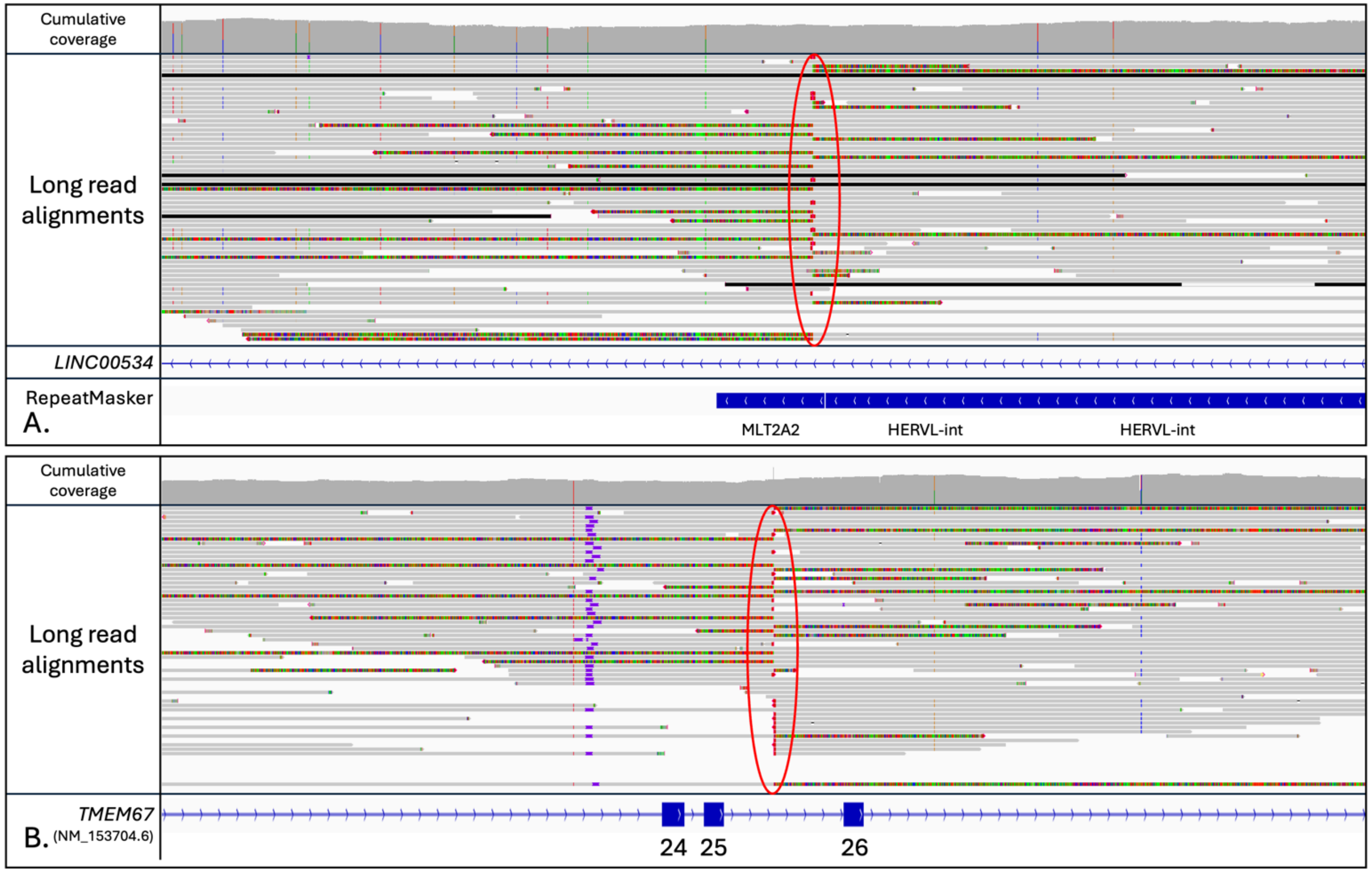

Figure 5.19: Long read sequence reads defining a putative structural variant that intersects the chromosome 8 target locus in paternal DNA sample. (A) Sequence reads that define the *LINC00534* breakpoint (red oval). **(B)** Sequence reads that define the *TMEM67* breakpoint (red oval). Sequence reads were aligned to build GRCh38 of the human reference genome. Exons for *TMEM67* are displayed as blue rectangles and numbered according to transcript NM_153704.6. The y-axis of cumulative coverage graphs is scaled 0-75x.

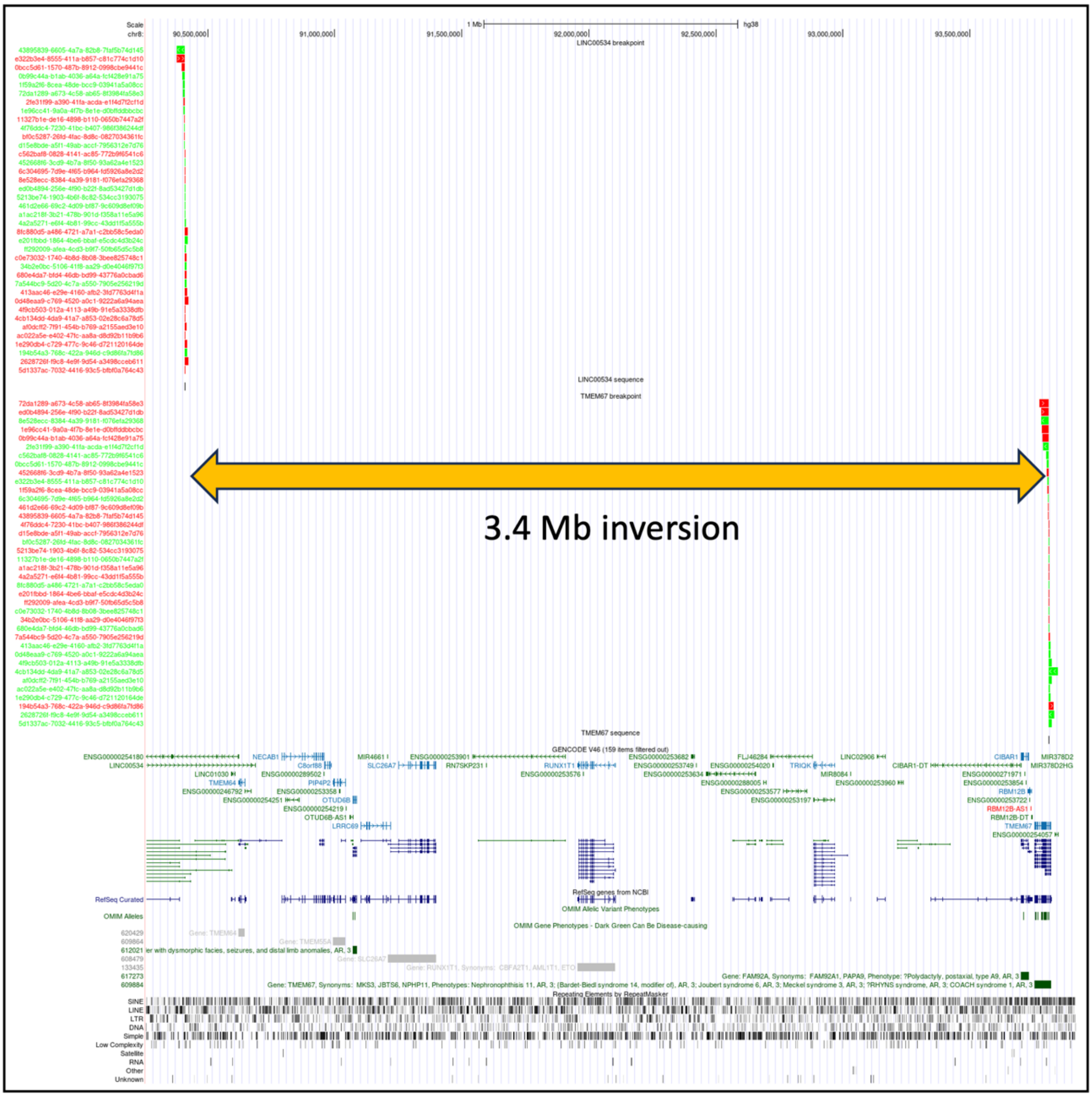

Figure 5.20: The coordinates of putative inversion spanning reads were resolved using BLAT then visualised using the UCSC Genome Browser. A 3.4 Mb inversion was identified. Genomic coordinates are displayed using build GRCh38.

The inversion was not detected in the original exome dataset due to the breakpoints occurring in intronic gene regions that were not targeted for analysis. Furthermore, as it was the paternal DNA that was analysed by long-read sequencing it was not clear whether the inversion had been inherited by the fetus. To inform this consideration a search for single nucleotide variants that were located in the vicinity of TMEM67 variants was performed. The non-reference nucleotide chr8:93782392T (c.1066-3T) was identified to be in *cis* with four inversion-spanning long-reads (Figure 5.21A). By re-analysing the exome sequencing data, it was evident that the fetus was heterozygous for the c.1066-3C>T variant. As the c.1066-3T allele tags the inversion-containing haplotype this indirect test was consistent with the fetus having a diagnosis of Meckel Gruber Syndrome.

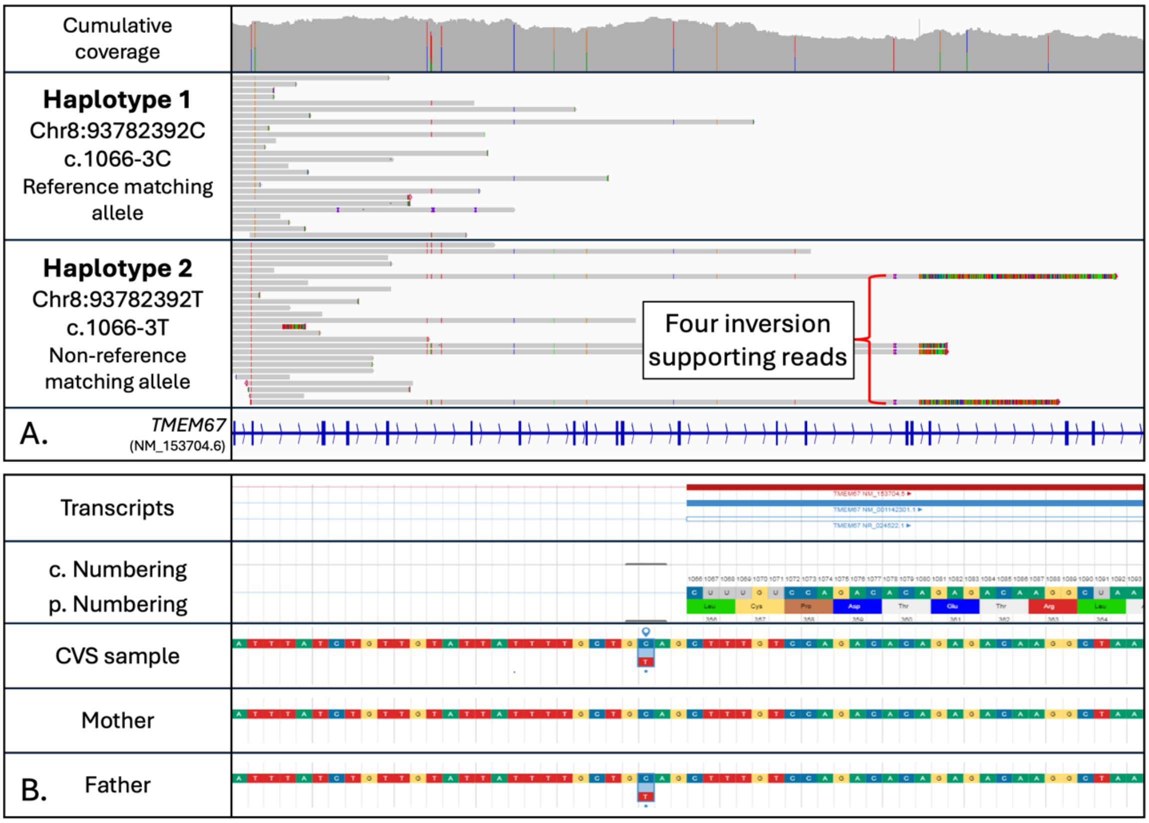

Figure 5.21: The paternal c.1066-3C>T variant, in cis with the putative TMEM67-intersecting inversion, is identified in the fetal CVS sample. (A) Both exonic variants and those that were in intronic sequence that was in close proximity to the exon were reviewed. The c.1066-3T variant was identified, which was in *cis* with four inversion spanning reads. **(B)** Exome sequencing performed on the fetal CVS sample revealed the c.1066-3T allele had been inherited from the father (providing indirect support that the fetus had inherited the inversion-containing allele).

PCR assays were optimised to allow Sanger sequencing of the c.1046T>C variant (Figure 5.22A) and the inversion breakpoints. Sanger sequencing of the PCR amplification products identified a 22 bp insertion in the *TMEM67* breakpoint (Figure 5.22B), in addition to an apparent 2 bp deletion at the *LINC00534* breakpoint (Figure 5.22C).

To enable a facile diagnostic test for detecting the inversion in extended family members a multiplex PCR assay was devised. The assays combined breakpoint-spanning primers with a third primer that amplified the wild type allele and could be resolved by agarose gel electrophoresis (Figure 5.23).

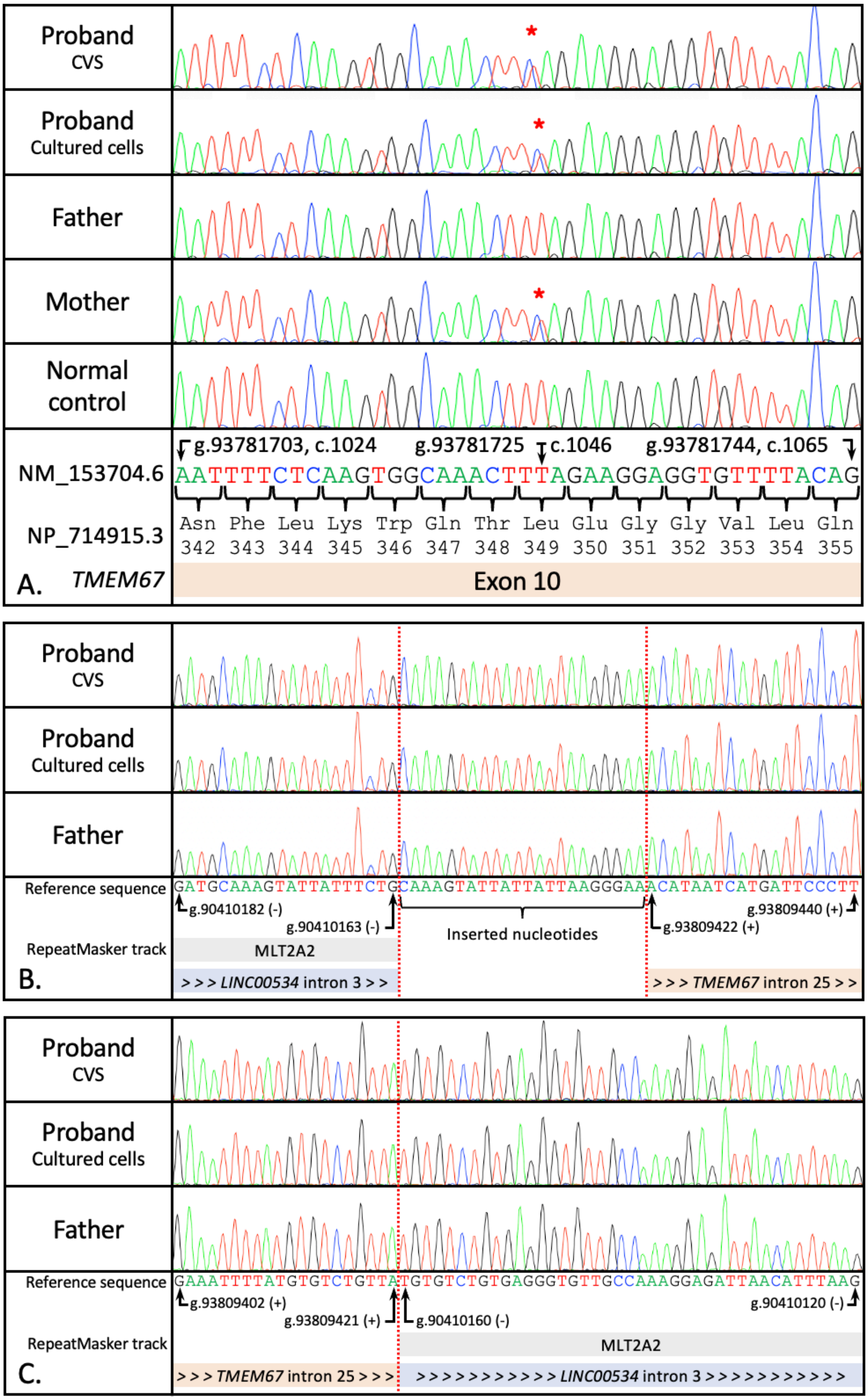

Figure 5.22: Sanger sequencing chromatograms verifying the c.1046T>C variant and enabling complete characterisation of the inversion breakpoints.

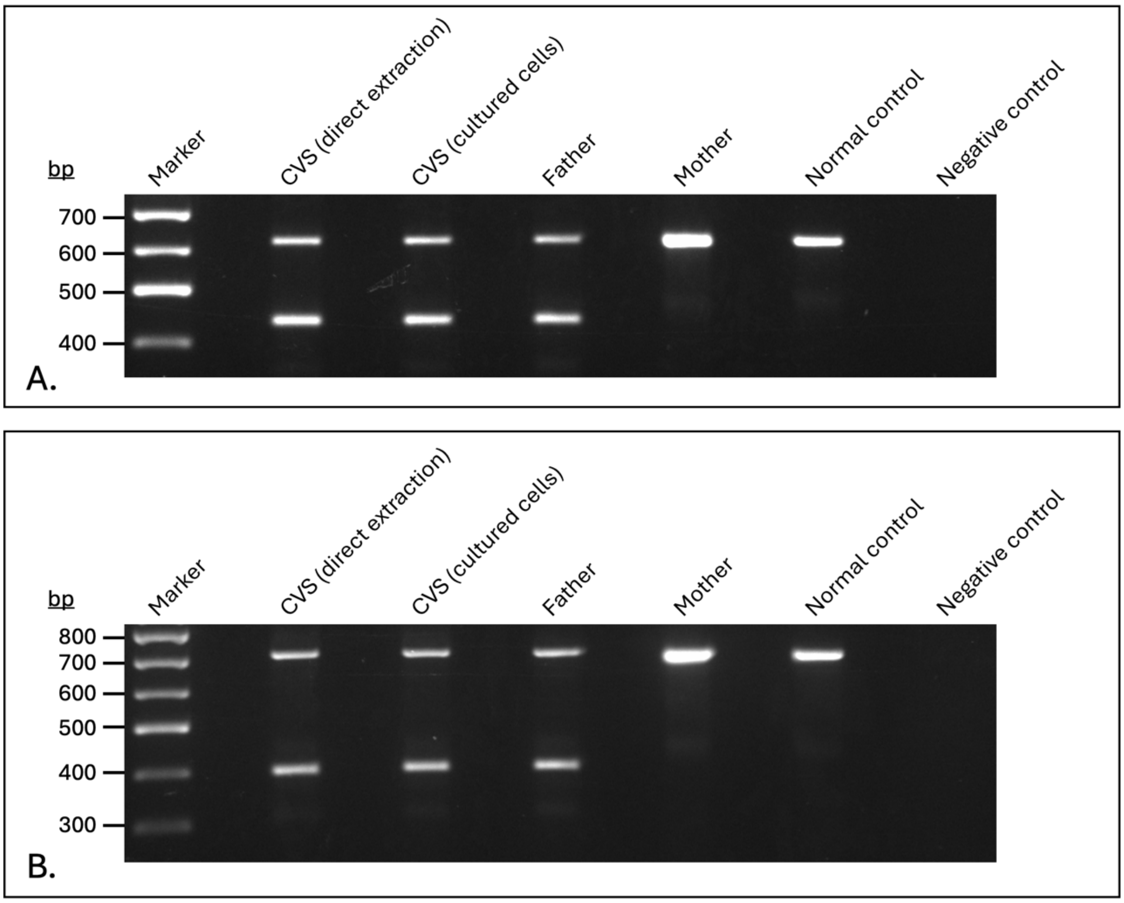

Figure 5.23: Breakpoint spanning PCR assay to verify the inversion-containing allele. (A) Amplification products spanning the *TMEM67* intersecting breakpoint are resolved to approximately 425 bp. The wild type allele is a 628 bp product (visible in all samples) **(B)** Amplification products spanning the *LINC00534* intersecting breakpoint are resolved to approximately 398 bp. The wild type allele is a 712 bp amplification product (visible in all samples). Both gels confirm the fetus has inherited inversion-containing allele from the father. The DNA ladder is the Norgen PCR Ranger marker.

5.5 Discussion

This chapter sought to use long-read whole genome sequencing to investigate patients in whom standard-of-care diagnostic testing had identified only a single heterozygous pathogenic mutation in an autosomal recessive disease gene with the aim of identifying the second pathogenic allele. Two exemplar cases as presented which demonstrate the capability of long-read WGS to deliver patient diagnoses when short-read sequencing does not.

Case 1

Using a custom gene panel test, this female patient had a single heterozygous variant, c.2808G>A (NM_014639.4), identified in exon 28 of *TTC37*, which was predicted to create a premature termination codon in the translated protein p.(Trp936*). No other candidate variants were detected in the coding region of the gene, nor in the immediate canonical splice sites. As the *TTC37* gene is associated with tricho-hepato-enteric syndrome (THES), a disorder which mirrors the patient phenotype of IUGR, woolly, brittle hair, and intestinal failure with congenital diarrhoea, this variant appeared likely to be one of two causative variants.

It has previously been reported by Bourgeois and colleagues that in a cohort of 96 patients with THES, around a quarter of the pathogenic alleles discovered were abnormal splicing variants in *TTC37* (Bourgeois *et al.,* 2018). This included a report of a cryptic splicing event (Hartley *et al.,* 2010b). It therefore seemed plausible that a splicing abnormality could be a cause of disease in this case and, given the canonical splice sites were normal in this proband, perhaps a deeper intronic variant could be present. *TTC37* is a large gene; 91 kb (hg19) with 43 exons. Twenty three of the 42 introns are longer than 1 kb, with 10 of them being over 3 kb (Bourgeois *et al.,* 2018). Therefore, while short-read WGS would potentially identify a deep intronic mutation, it would not be possible to phase a variant using short reads. The patient was therefore a good candidate for implementing long-read technology to try and resolve the case.

The DNA sample was first assessed for fragment size distribution using the Femto Pulse, and in this case both a fragmentation step and short-read elimination treatment were required. A library was prepared for sequencing on the PromethION. While 1 ug of input DNA was used for this preparation, a relatively low yield of 163.8 ng was returned. There are a number of possibilities for this. Firstly, the assumed input quantity for the preparation may have been inaccurate due to pipetting error or issues with quantification. As the subsequent library preparation based on the same volume of input DNA had a return yield of 422.4 ng, the quantification of input material was most likely correct. A more likely explanation for the low yield initially, is the two AMPure beads clean-ups performed during the library preparation. The success of any SPRI bead-based technique is reliant on thorough mixing of the DNA with the beads to allow optimal binding. Conversely, a key consideration when performing long-read sequencing preparation is ensuring that no vortexing or excessive pipette mixing of the sample is performed. Gentle inversion of tubes, flick-mixing and/or rotation of the sample/bead mixture is performed. It is plausible that in the first instance, the mixture was not sufficiently resuspended at some point meaning DNA was lost to sub-optimal bead binding or elution.

Nuclease flushing of the flow cell, and reloading of the sample, was utilised to maximise data yield from the sequencing run, with a resultant yield of 104.44 Gb and genome coverage of approximately 30x achieved. Alignment of reads and analysis of the entirety of *TTC37* identified three heterozygous variants, one of which was the previously detected c.2808G>A. The reads were phased with respect to their allelic composition at this site, demonstrating a key advantage of long read technology. In doing so, a single candidate heterozygous variant in the intron between exons 25 and 26 was identified.

While interpreting the pathogenicity of intronic variants remains complex (Walker *et al.,* 2023), bioinformatics tools with ever-greater sensitivity and specificity continue to emerge. Exemplar tools include SpliceAI (Jaganathan *et al.,* 2019) and Pangolin (Zeng & Li, 2022), these aim to predict the impact of intronic variants on gene function. Analysis of the identified heterozygous candidate variant, c.2634+679A>G, using SpliceAI suggested the possible pathogenic nature of the variant with a probability score of 0.98 for a cryptic donor site being generated by this variant. Nevertheless, best practice guidelines for variant classification (Durkie *et al.,* 2024) state that *in silico* splicing predictions are only indicative of pathogenicity and further supporting evidence, such as RNA studies, should be obtained to comprehensively characterise the variant. Patient RNA was therefore extracted from a fresh peripheral blood sample and used to perform short read transcriptome sequencing, in order to verify the suspected aberrant splicing. The identification of a novel exon occurring between exons 25 and 26, with a cumulative read depth of approximately half that of other exons, suggested that the variant did induce aberrant splicing of the affected allele, satisfying the basic requirements of functional evidence of pathogenicity.

As this work was primarily conducted to assess the utility of long read sequencing, it was decided to take the RNA sample from the patient forward for long read cDNA sequencing, with the aim of determining phase at the level of transcription. A long-read cDNA library was prepared and sequenced on the PromethION, from which a single read showed both the novel exon and the reference base at the position of the known pathogenic variant, demonstrating that the novel exon was occurring on the alternate haplotype. With approximately 85 million long reads generated, and *TTC37* being expressed at a level of approximately 4.4 transcripts per million in peripheral blood according the UCSC Genome Browser, one would anticipate a higher number of reads covering this isoform. There could be a number of reasons why only a single read was obtained. The longest transcript of the gene is approximately 5.7 kb (da Silva Franco *et al.,* 2025). The N50 for this run was 1.02 kb, whereas suggesting that lengths of the cDNA obtained were not sufficiently long to achieve full length transcripts. This could be related to the quality of the RNA used, as the RIN^{e} value of 7.9 given by the TapeStation suggests some degradation of the sample, meaning the longest transcripts may not be represented in their entirety. The PCR conditions of the cDNA library preparation may also have been a factor; 14 cycles were used based on the recommendations in the manufacturers protocol for the kit. This recommendation is based on using the lowest number of cycles while still generating the 50 fmol of library required to load the flow cell, however, Oxford Nanopore do highlight that more PCR cycles leads to bias towards shorter reads (Oxford Nanopore Technologies, no date a), and the fewest PCR cycles should be used to achieve the required library yield for sequencing. The yield achieved for this sample was 57.6 ng, equating to approximately 93.5 fmol (based on an average cDNA size of 1 kb), and a little over half that amount was used for loading the flow cell, meaning there was some excess material in this case. Future work using this technique for analysis of longer transcripts could include optimisation to reduce the PCR cycles for this step.

While there are technical challenges associated with the assay, this case does demonstrate the potential for a long-read solution for both whole genome and transcriptome sequencing to resolve cases of rare recessive disease, which offers the benefit of phasing variants to aid interpretation.

Case 2

The pregnancy in this case presented at ultrasound with characteristic features associated with Meckel Gruber syndrome. By using exome sequencing the discovery of a single heterozygous variant in *TMEM67* was consistent with an incomplete diagnosis of this disorder. The second variant remained undetected by short read sequencing, and so the case was a good candidate for long read whole genome sequencing. The material obtained from the fetus was a chorionic villus sample, therefore the quantity of DNA available was limited and insufficient for long read library preparation. As trio analysis had been performed as part of the prenatal exome pathway, the *TMEM67* variant had also been detected in the mother. The father was thereby the obvious candidate for long read sequencing as the presumed carrier of the second, as yet undetermined variant.

The DNA sample from the father was assessed using a Femto Pulse. In contrast to the initial QC of the DNA for Case 1, the sample had a more refined size distribution, with the majority of fragments measured between 10 kb and 50 kb. A small distribution of lower molecular weight fragments was seen (Figure 5.16), however these were predominantly below the 1.3 kb marker. As a Long Fragment Buffer wash is used during the library preparation clean-up process, this should remove fragments below 3 kb. The sample was therefore taken forward for library preparation and sequencing without any additional fragmentation or short read elimination being performed. The subsequent estimated read length distribution for the run showed a large proportion of the resulting sequencing reads being below 7.5 kb (Figure 5.18) compared to the previous case (Figure 5.4) which had been treated with SRE. This could suggest the Long Fragment Buffer did not effectively remove all DNA fragments below 3 kb. At this stage the DNA and bead mixture is quite viscous and therefore mixing the solution until homogenous is critical to efficient size selection. There is also the possibility that the Femto Pulse measurement may not be a true reflection of the exact DNA sizing, and the fragments may have been of higher molecular weight than measured. While this skew towards shorter reads did not have a negative impact on the analysis in this case, it should be taken into consideration with future samples. For some analysis, such as the previously discussed PMS2 gene for example, maximising the length of the reads is key to assay success, and therefore utilising short read elimination treatment even in samples with a low proportion of fragments below 10 kb is a worthwhile endeavour, as the difference in read distribution between these two samples demonstrates.

Review of long read whole genome sequencing data at the target gene *TMEM67* was able to identify a 3.4 Mb inversion intersecting this locus and *LINC00534*. As with Case 1, haplotype phasing was used to support the finding. Long reads containing the inverted sequence were demonstrated to occur in *cis* with a c.1066-3T variant in the father's sample. Following a reanalysis of the exome data, the c.1066-3T variant was detected in both the fetus and father, but not the mother, providing indirect supporting evidence of the fetus being a carrier of the inversion. It was subsequently possible to design primers to verify the inversion breakpoints by Sanger sequencing and confirm the presence of both pathogenic variants in the fetus. This showed the rearrangement occurring at *TMEM67* intron 25 and *LINC00534* intron 3, with a 22bp insertion occurring at the *TMEM67* breakpoint and the apparent deletion of 2bp of the

LINC00534 intron at that breakpoint. A pair of multiplex assays also allowed agarose gel visualisation of carrier status of the wild type and alternate alleles in the family.

This case is an example of large structural variation identified by nanopore sequencing which would not otherwise have been easily detected by short ready sequencing methods. As a prenatal case this work highlights the potential use of this technology to aid rapid diagnosis in urgent referral pathways. The work described here was however facilitated by a priori knowledge of the parents' carrier status of the known pathogenic allele, allowing use of the paternal DNA rather than precious fetal material. At present, the high DNA input requirements of nanopore sequencing are likely to be limiting factor in being able to deploy this technology directly to prenatal samples. There is the possibility to utilise DNA obtained from cell cultured CVS samples, but this would add additional processing time to the testing pathway.

The result obtained here via long read sequencing allows appropriate genetic counselling to be offered to the family based on their known carrier status and will enable prenatal testing in future pregnancies should this be desired.

5.6 Summary

The two cases reported here demonstrate the utility of long read sequencing in rare recessive disease patients where a single pathogenic allele has been detected in a gene that is consistent with the phenotype. Although intronic variants can be identified from short-read sequencing datasets, their interpretation from long-read whole genome sequencing is aided by an ability to create phased haplotypes. Indeed, haplotype phasing has been shown to be a valuable diagnostic tool in both cases. The detection of a large structural variant is something which would be difficult to achieve by short read WGS. The results here therefore support roll-out of this technology to selected patients where standard of care testing has provided a partial diagnosis of recessive disease.

6. Closing Discussion

Short-read next-generation sequencing, while revolutionary in terms of expanding the scope and scale of genomic screening, has some inherent limitations which may impede analyses of certain referral types. The aim of this research was to assess the utility of emerging technologies in addressing these current challenges to improve diagnostic workflows and outcomes for patients.

Aneuploidy screening from whole genome amplified material

Initially this work was focussed on single cell sequencing. At the time, single cell array CGH had been reported as a promising tool for aneuploidy screening of embryo biopsies in reproductive medicine. Due to an established local workflow for copy number variation sequencing using NGS, as an alternative to array CGH in constitutional paediatric cases, this seemed like a logical starting point to approach this method. Aligning to the aims of the project, the principal challenge when sequencing single cells using NGS is the requirement for nanogram quantities of DNA. Whole genome amplification is therefore required to enable sequencing, and the first section of this research was a preliminary study of a simple, cost-effective method for WGA, and PCR-based methods which could be used to assess the performance of the assay with a view to optimisation and improvement. The presented results demonstrate that it was challenging to gain insight on the overall performance of whole genome amplification without performing whole genome sequencing. In part, this was due to the limited number of markers utilised; perhaps a larger scale data set may have aided this approach. Nevertheless, using CNVseq data as an overall assessment of the multiple displacement amplification WGA method, a large amount of development would be required to improve the performance of the assay. While there are many optimisations reported, including alterations to temperature, incubation time, reaction volume (Marcy *et al.,* 2007), and use of emulsion droplets for amplification (Rhee *et al.,* 2016), it became apparent over the course of the investigation that opinions on the efficacy of aneuploidy screening in embryo biopsies, to improve pregnancy outcomes, were mixed (Griffin, 2022). Furthermore, the restructuring of genomics testing in the NHS meant that this research no longer fell within the remit of our group, and the preliminary study became a natural endpoint for this work. To date, this type of pre-implantation aneuploidy screening is not funded by the NHS but is offered by some private fertility providers.

While the utility of aneuploidy screening in reproductive medicine remains contentious, from a technology perspective, multiple displacement amplification continues to be adopted in a variety of research fields where source material is limited (Ospino *et al.,* 2024), and the use of long-read sequencing will no doubt be the next step for those utilising MDA for sequencing purposes. Reports of nanopore sequencing of MDA products (Agyabeng-Dadzie *et al.,* 2025) describe computational removal of concatemer sequences in order generate high quality microbial genome assembles, however Oxford Nanopore Technologies do acknowledge that the structure of MDA products leads to reduced sequencing yields and quality, with pore blockages due to the concatemerized, branching structure of the products (Oxford Nanopore Technologies, no date b). Use of an endonuclease digestion to remove branch structures prior to sequencing is advised. Notably, preimplantation genetic studies using an alternative WGA method have shown the benefit of nanopore sequencing in this field. (Madritsch *et al.,* 2024) and one could envisage that the low capital investment and portability of devices such the Oxford Nanopore MinION sequencer could make the technology easily accessible for reproductive medicine units to have on site.

In summary, this chapter of the thesis provided valuable insight into amplification bias issues which are inherent to this technique but can also be seen in NGS technology as a whole. As more and more library preparation kits come to market with ever decreasing sample input requirements, it is important to consider the implications for bias across the genome. Indeed, this is why PCR-free library preparation is the preferred method for whole genome sequencing assays (G. Zhou *et al.,* 2022), and why the use of native DNA sequencing technologies which do not introduce an amplification bias, such as Oxford Nanopore, is promising for the future of WGS.

Targeted analysis of homologous loci

The research moved on to investigate challenging loci which impede patient diagnosis due to limitations of assay design. Targeted hybridisation/capture assays have become common place in genomic diagnostics, achieving a balance between sequencing a large number of medically relevant loci simultaneously whilst costing significantly less than sequencing the entire genome. The nature of these assays is such that effective enrichment relies on target sequence being entirely unique within the genome. While this approach enables capture of the majority of coding exons in the genome, problems arise when genes with a number of other homologous loci are targeted. Pseudogenes are the primary example of this scenario and are particularly problematic when the gene/pseudogene homology is extensive, such is the case with the *PMS2* gene and *PMS2CL* pseudogene. To confirm a variant detected in an area with high homology, it must be linked to a unique region of the gene. In the case of *PMS2*, previously this would only have been possible by long range PCR, something which adds cost and complexity to the testing pathway.

Linked-read sequencing was able to overcome this issue by enabling the enriched short sequencing reads to be reconstructed into their originating DNA molecule, allowing the assessment of the variants detected in the context of the wider locus, and distinguishing between gene and pseudogene by anchoring linked reads in a unique locus. However, with the emergence of long read sequencing, one could propose that this may be an alternative route by which the detection of variants in *PMS2* and similar loci could be achieved. From the linked-read cohort, there was a sample failure due to low input DNA quality, suggesting that sample degradation would be a factor to consider when using this approach. While the same is true of samples used for long read sequencing, an advantage of this approach is that the DNA is sequenced natively, generating data reads which are directly representative of the original DNA molecule. With linked-read technology, the low DNA mass inputted to the assay coupled with multiple rounds of amplification the sample undergoes during library preparation and sequencing, means the final sequence data is an extrapolation of the originating DNA strands. The output data of the assay is therefore more severely affected by DNA degradation due to fewer molecules representing the target locus being included to begin with. By using direct DNA sequencing by nanopore, a sample such as the one reported here, would generate a proportion of long reads in addition to the many shorter reads resulting from DNA degradation. Furthermore, there is the option to size select out the shorter DNA fragments prior to sequencing. It would be reasonable to assume that the chance of sampling intact multi kilobase fragments covering the region of interest would be higher with this method. However, in order to reduce cost and obtain the sequencing coverage required to accurately characterise variants, a targeted approach may be preferable.

There are reported hybridisation capture assays which are compatible with long read sequencing, including Agilent (Agilent Technologies, no date) and Twist (Twist Bioscience, no date). These methods report a target fragment size of ~5 kb (Agilent) up to 20 kb (Twist). For some applications this could be sufficient to achieve reads covering the region of interest, but in the context of *PMS2*, would not be long enough to achieve the 25 kb red required to anchor in the unique region. Adaptive sampling using Oxford Nanopore sequencing also offers an alternative targeted sequencing approach. This approach could be applied to generate a targeted gene panel similar to the one currently deployed for short read sequencing, with the added benefit of being able to identify and distinguish between *PMS2* and *PMS2CL*. By sequencing DNA natively, one would also anticipate a reduction in the PCR biases inherent to hybridisation capture assays and the ability to sequence into genomics regions which are not conducive to *in vitro* enrichment methods, such as repetitive sequences and GC-rich loci.

As adaptive sampling relies on partial sequencing and rejection of off-target sequences in order to enrich for target loci, preparing a high-quality library to maximise pore occupancy on the flow cell is key to ensuring the maximum number of molecules are sequencing at any one time. From our own experience, maintaining high pore occupancy and data yield throughout sequencing is contingent on DNA fragments being sufficiently fragmented to prevent pore blockages occurring. That being said, in order for the sequencing reads generated to be useful, they must be sufficiently long to encapsulate the regions of interest. In the case of *PMS2*, reads in the region of 25 kb are required in order capture the unique exon 6-8 region along with the 3' gene regions which are homologous to *PMS2CL*. This size specifically is not likely to result in major pore blocking, however the challenge is shearing DNA as closely to this size as possible. Covaris G-tubes are one option to fragment to multi-kilobase sizing, but with this approach the size distribution can be variable and often quite broad. In ensuring the complete fragmentation of the highest molecular weight material which leads to pore blocking, the overall size profile becomes smaller. An alternative method to achieve more targeted, precise shearing is the Megaruptor series of instruments (Hologic, USA), an automated shearing platform which employs a "Hydropore" device through which DNA is forced at controlled speeds in order to dictate the fragment size generated.

While long read sequencing may have the potential to aid variant classification in some homologous loci, challenges remain around sample preparation for this workflow to achieve maximum read length and yield simultaneously. Furthermore, the nanopore sequencing devices are, in comparison to Illumina sequencers, quite laborious to operate and remain limiting when it comes to sample throughput. Therefore, the promise of a short-read solution such as linked reads remains appealing for gene panel targets such as *PMS2* in terms of fitting with existing infrastructure in high throughput genomics laboratories. It remains to be seen whether alternative linked reads kits such as TELL-Seq or new strategies such as the Illumina constellation mapping will help short read sequencing remain competitive in this arena as long-read sequencing solutions continue to evolve.

Long read sequencing in rare disease

It is acknowledged that the high throughput targeted short read assays used for most rare disease genomic tests do not detect the full spectrum of pathogenic variation. As sequencing costs decrease, whole genome short read sequencing is now becoming more commonplace as an alternative. While this opens up the scope of the rare disease test and allows detection of novel intronic variation, it does not offer the additional information required to aid interpretation of recessive variants, such as generating parental haplotypes. Furthermore, larger structural rearrangements remain refractory to short read sequencing analysis.

Here we have demonstrated that for selected cases long read whole genome sequencing using the Oxford Nanopore PromethION can resolve cases of rare recessive disease where a second pathogenic allele has not been detected by short read sequencing. While the two cases described here were successful, there remain some challenges associated with the use of this technology, and these should be considered as part of future development of rare disease services.

As has been demonstrated with this research, DNA sample quality and size distribution are critical to achieving optimal flow cell pore occupancy, sequencing read lengths and data yield. The methods utilised here for DNA fragmentation and size selection were manual methods of sample preparation and would be extremely laborious to undertake in the laboratory at scale. While there are reports of automation solutions for the subsequent library preparation steps undertaken (Tecan, no date), the loading process for the PromethION itself remains entirely manual and the hands-on time required is exacerbated by the need to perform daily nuclease washes and sample reloads to achieve maximum data yield. Oxford Nanopore Technologies have their own automated library preparation platform in development, which will have sequencing set-up included, but currently plans are centred around the lower capacity P2 Solo version of the PromethION. With local throughput of rare disease patients for targeted and WGS approaches equating to hundreds of samples per week, the PromethION is not currently able to meet the service demands of first-line testing.

Nevertheless, when the technology is applied in appropriate contexts, it can yield results for patients who would otherwise be without a conclusive diagnosis, and the benefit of this cannot be understated. In recessive rare disease these results enable appropriate counselling and support to be given and can have much wider implications beyond a single patient; knowledge of the mutation carrier status can allow cascade testing to other family members and can inform reproductive choices. Therefore, although long read nanopore sequencers may not yet be deployed as a wholesale short read sequencing alternative, there is strong evidence for use as a reflex test to further investigate patients with a single recessive variant with a strong phenotypic association.

There are, however, some patient cohorts for which long read sequencing is already looking set to challenge the dominance of previous standard of care technologies. Methylation profiling of central nervous system tumours using Oxford Nanopore sequencing looks set to significantly alter the patient care pathway. With reports of the ability to classify tumours while the patient is still on the operating table, (Vermeulen *et al.,* 2023) one can envisage that the current approach of testing by methylation array, which can take weeks to return results, may eventually become redundant. Promising results have also been obtained in rapid detection of structural variants in oncology samples (Elrick *et al.,* 2025).

While the work here has focussed on the Oxford Nanopore Technologies PromethION sequencing platform, there are other vendors which may become competitors in the long-read sequencing field. There have been recent publications demonstrating improved diagnostic yield from the PacBio Revio when sequencing rare disease patients. Höps and colleagues were able to demonstrate the efficacy of the Revio in detecting variants which are challenging to elucidate via short read sequencing, and which had previously been identified using orthogonal technologies (Höps *et al.,* 2025). Furthermore, Steyaert and colleagues were able to detect novel variants in previously unresolved patient cases (Steyaert *et al.,* 2024). With reported read lengths of up to 15 kb in these studies, this platform also enables detection of structural variation beyond the capability of short read sequencing, although AG repeat expansions remain a challenge, as do structural breakpoints located in repetitive (Höps *et al.,* 2025). This suggests an advantage of the native DNA sequencing employed by Oxford Nanopore Technologies platforms, over the HiFi sequencing used in PacBio platforms which remains reliant on polymerase amplification of DNA sequence. As well as the PacBio Revio, Roche have recently announced their sequencing by expansion (SBX) technology (Kokoris *et al.,* 2025); template nucleotides are separated into an expanded molecule, or “Xpandomer”, which is then fed through a nanopore sequencer. While only generating a maximum read length of 1000 bp, this spatial separation of target nucleotides is reported to improve accuracy and throughput compared to other nanopore sequencing approaches, with claims of sequencing 7 genomes at 30x coverage per hour and 99.8 % accuracy in SNV calling reported.

While the performance of SBX technology is yet to be scrutinised in the wider field, Oxford Nanopore sequencing is already making a real-world difference to patient diagnosis. In particular, the work of Matt Loose and team at the University of Nottingham looks set to help transform the tumour classification pathway in the NHS, with test results that were previously taking weeks now possible in days or merely hours thanks to nanopore technology (Deacon *et al.,* 2025). Indeed, in our own laboratory we are seeing the benefits of this technology for patients, with nanopore sequencing utilised to resolve a number of published cases (Watson, Crinnion, *et al.,* 2020b; Watson *et al.,* 2022; McClinton, Crinnion, *et al.,* 2023).

7. Future work and conclusions.

Long read sequencing appears to be on the cusp of adoption into routine diagnostic care in the NHS. Whether long read sequencing could eventually become the frontline test remains to be seen, and an evolution of the technology akin to that which occurred for short read sequencers would be required in order to meet the scale of testing currently undertaken by short read analysis.

The work reported here lends support to the immediate utilisation of this technology in unresolved cases of rare recessive disease and also in cases where larger or more complex structural variation is suspected. It is also our ambition to try to resolve *PMS2* variation using long read sequencing, although there are technical limitations to be overcome due to the size of this gene. Should this not be possible, then alternative linked read methods may be pursued.

While genomics has been transformed by various iterations of DNA sequencing technology over the last thirty years, it is evident that there remain advancements to be made in improving diagnostic yield in genetic testing. It is therefore important that we continue to assess the capability of new technologies. In doing so we not only help our individual patients, but in cases such as *PMS2* testing, we can improve screening and disease prevention rates, something which is of benefit to the NHS as a whole.

8. References

10X Genomics (no date) *https://assets.ctfassets.net/an68im79xiti/6ceYcRzVAc6MaSMeyO0akE/4d9f269143be9e1750a415e1d5aa6762/CG00044_10X_Techical_Note_LinkedReads.pdf, 2019*.

Abdelhamed, Z.A. *et al.* (2015) 'The Meckel-Gruber syndrome protein TMEM67 controls basal body positioning and epithelial branching morphogenesis in mice via the non-canonical Wnt pathway.', *Disease models & mechanisms*, 8(6), pp. 527–41. Available at: https://doi.org/10.1242/dmm.019083.

Agilent Technologies (no date) 'an-long-read-sureselect-xt-hs2-5994-7612en-agilent'.

Agyabeng-Dadzie, F. *et al.* (2025) 'Evaluating the Benefits and Limits of Multiple Displacement Amplification With Whole-Genome Oxford Nanopore Sequencing.', *Molecular ecology resources*, 25(6), p. e14094. Available at: https://doi.org/10.1111/1755-0998.14094.

Ahn, M.-J. *et al.* (2019) 'Osimertinib in patients with T790M mutation-positive, advanced non-small cell lung cancer: Long-term follow-up from a pooled analysis of 2 phase 2 studies.', *Cancer*, 125(6), pp. 892–901. Available at: https://doi.org/10.1002/cncr.31891.

Arneson, N. *et al.* (2008) 'Whole-genome amplification by degenerate oligonucleotide primed PCR (DOP-PCR)', *Cold Spring Harbor Protocols*, 3(1). Available at: https://doi.org/10.1101/pdb.prot4919.

Bentley, D.R. *et al.* (2008) 'Accurate whole human genome sequencing using reversible terminator chemistry', *Nature*, 456(7218), pp. 53–59. Available at: https://doi.org/10.1038/nature07517.

Blainey, P.C. and Quake, S.R. (2011) 'Digital MDA for enumeration of total nucleic acid contamination.', *Nucleic acids research*, 39(4), p. e19. Available at: https://doi.org/10.1093/nar/gkq1074.

Bouras, A. *et al.* (2024) ‘PMS2 or PMS2CL? Characterization of variants detected in the 3′ of the PMS2 gene’, *Genes Chromosomes and Cancer*, 63(1). Available at: https://doi.org/10.1002/gcc.23193.

Bourgeois, P. *et al.* (2018) ‘Tricho-Hepato-Enteric Syndrome mutation update: Mutations spectrum of TTC37 and SKIV2L, clinical analysis and future prospects’, *Human Mutation*, 39(6), pp. 774–789. Available at: https://doi.org/https://doi.org/10.1002/humu.23418.

Burke, W. *et al.* (2022) ‘The Challenge of Genetic Variants of Uncertain Clinical Significance : A Narrative Review.’, *Annals of internal medicine*, 175(7), pp. 994–1000. Available at: https://doi.org/10.7326/M21-4109.

Burn, J. *et al.* (2011) ‘Long-term effect of aspirin on cancer risk in carriers of hereditary colorectal cancer: an analysis from the CAPP2 randomised controlled trial.’, *Lancet (London, England)*, 378(9809), pp. 2081–7. Available at: https://doi.org/10.1016/S0140-6736(11)61049-0.

Carr, I.M. *et al.* (2013) ‘Simple and efficient identification of rare recessive pathologically important sequence variants from next generation exome sequence data.’, *Human mutation*, 34(7), pp. 945–52. Available at: https://doi.org/10.1002/humu.22322.

Carvalho, F. *et al.* (2020) ‘ESHRE PGT Consortium good practice recommendations for the organisation of PGT†’, *Human Reproduction Open*, 2020(3). Available at: https://doi.org/10.1093/hropen/hoaa021.

Chen, E. *et al.* (2023) ‘Rates and Classification of Variants of Uncertain Significance in Hereditary Disease Genetic Testing.’, *JAMA network open*, 6(10), p. e2339571. Available at: https://doi.org/10.1001/jamanetworkopen.2023.39571.

Chen, S. *et al.* (2024) ‘A genomic mutational constraint map using variation in 76,156 human genomes.’, *Nature*, 625(7993), pp. 92–100. Available at: https://doi.org/10.1038/s41586-023-06045-0.

Chen, Z. *et al.* (2020) ‘Ultralow-input single-tube linked-read library method enables short-read second-generation sequencing systems to routinely generate highly accurate and economical long-range sequencing information’, *Genome*

Research, 30(6), pp. 898–909. Available at: https://doi.org/10.1101/gr.260380.119.

CHUGH, K. (2024) 'Transforming "Junk" DNA into Cancer Warriors: The Role of Pseudogenes in Hepatocellular Carcinoma', *Cancer Diagnosis & Prognosis*, 4(3), pp. 214–222. Available at: https://doi.org/10.21873/cdp.10311.

Clark, M.J. *et al.* (2011) 'Performance comparison of exome DNA sequencing technologies.', *Nature biotechnology*, 29(10), pp. 908–14. Available at: https://doi.org/10.1038/nbt.1975.

Coonen, E. *et al.* (2020) 'ESHRE PGT Consortium good practice recommendations for the detection of structural and numerical chromosomal aberrations†', *Human Reproduction Open*, 2020(3). Available at: https://doi.org/10.1093/hropen/hoaa017.

Crick, F. (1958) 'Periodical. Crick, Francis. "On Protein Synthesis." The Symposia of the Society for Experimental Biology 12, (1958): 138-163. Article. 13 Images.. The Symposia of the Society for Experimental Biology'.

Cummings, B.B. *et al.* (2017) 'Improving genetic diagnosis in Mendelian disease with transcriptome sequencing.', *Science translational medicine*, 9(386). Available at: https://doi.org/10.1126/scitranslmed.aal5209.

Deacon, S. *et al.* (2025) 'ROBIN: A unified nanopore-based assay integrating intraoperative methylome classification and next-day comprehensive profiling for ultra-rapid tumor diagnosis.', *Neuro-oncology*, 27(8), pp. 2035–2046. Available at: https://doi.org/10.1093/neuonc/noaf103.

Dean, F.B. *et al.* (2002) *Comprehensive human genome amplification using multiple displacement amplification*. Available at: www.pnas.orgcgidoi10.1073pnas.082089499.

DeAngelis, M.M., Wang, D.G. and Hawkins, T.L. (1995) 'Solid-phase reversible immobilization for the isolation of PCR products', *Nucleic Acids Research*, 23(22), pp. 4742–4743. Available at: https://doi.org/10.1093/nar/23.22.4742.

Delhanty, J.D.A. (1994) 'Preimplantation diagnosis', *Prenatal Diagnosis*, 14(13), pp. 1217–1227. Available at: https://doi.org/https://doi.org/10.1002/pd.1970141307.

Durkie, M. *et al.* (2024) *ACGS Best Practice Guidelines for Variant Classification in Rare Disease 2024*, *Royal Devon University Healthcare NHS Foundation Trust, Exeter, EX2 5DW. 5. Division of Genetics and Epidemiology*. Available at: http://gnomad.broadinstitute.org/.

Ebbert, M.T.W. *et al.* (2019) 'Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight', *Genome Biology*, 20(1), p. 97. Available at: https://doi.org/10.1186/s13059-019-1707-2.

Edwards, P. and Monahan, K.J. (2022) 'Diagnosis and management of Lynch syndrome', *Frontline Gastroenterology*. Available at: https://doi.org/10.1136/flgastro-2022-102123.

Ellingford, J.M. *et al.* (2022) 'Recommendations for clinical interpretation of variants found in non-coding regions of the genome.', *Genome medicine*, 14(1), p. 73. Available at: https://doi.org/10.1186/s13073-022-01073-3.

Elrick, H. *et al.* (2025) 'SAVANA: reliable analysis of somatic structural variants and copy number aberrations using long-read sequencing', *Nature Methods* [Preprint]. Available at: https://doi.org/10.1038/s41592-025-02708-0.

ESHRE Guideline Group on RPL *et al.* (2018) 'ESHRE guideline: recurrent pregnancy loss.', *Human reproduction open*, 2018(2), p. hoy004. Available at: https://doi.org/10.1093/hropen/hoy004.

Esteban, J.A., Salas, M. and Blanco, L. (1993) 'Fidelity of phi 29 DNA polymerase. Comparison between protein-primed initiation and DNA polymerization.', *The Journal of biological chemistry*, 268(4), pp. 2719–26.

Evans, D.G. *et al.* (2016) 'Comprehensive RNA Analysis of the NF1 Gene in Classically Affected NF1 Affected Individuals Meeting NIH Criteria has High Sensitivity and Mutation Negative Testing is Reassuring in Isolated Cases With Pigmentary Features Only.', *EBioMedicine*, 7, pp. 212–20. Available at: https://doi.org/10.1016/j.ebiom.2016.04.005.

Fang, H. *et al.* (2014) 'Reducing INDEL calling errors in whole genome and exome sequencing data.', *Genome medicine*, 6(10), p. 89. Available at: https://doi.org/10.1186/s13073-014-0089-z.

Ferguson-Smith, M.A. (2015) ‘History and evolution of cytogenetics.’, *Molecular cytogenetics*, 8, p. 19. Available at: https://doi.org/10.1186/s13039-015-0125-8.

Fleming, A. *et al.* (2024) ‘Combined approaches, including long-read sequencing, address the diagnostic challenge of HYDIN in primary ciliary dyskinesia.’, *European journal of human genetics : EJHG*, 32(9), pp. 1074–1085. Available at: https://doi.org/10.1038/s41431-024-01599-7.

FORD, C.E. *et al.* (1959) ‘A sex-chromosome anomaly in a case of gonadal dysgenesis (Turner’s syndrome).’, *Lancet (London, England)*, 1(7075), pp. 711–3. Available at: https://doi.org/10.1016/s0140-6736(59)91893-8.

Ganster, C. *et al.* (2010) ‘Functional PMS2 hybrid alleles containing a pseudogene-specific missense variant trace back to a single ancient intrachromosomal recombination event’, *Human Mutation*, 31(5), pp. 552–560. Available at: https://doi.org/https://doi.org/10.1002/humu.21223.

Geraedts, J. and Sermon, K. (no date) ‘Preimplantation genetic screening 2.0: the theory’. Available at: https://doi.org/10.1093/molehr/gaw033.

Gillooly, J.F., Hein, A. and Damiani, R. (2015) ‘Nuclear DNA content varies with cell size across human cell types’, *Cold Spring Harbor Perspectives in Biology*, 7(7), pp. 1–27. Available at: https://doi.org/10.1101/cshperspect.a019091.

Glassing, A. *et al.* (2016) ‘Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples’, *Gut Pathogens*, 8(1). Available at: https://doi.org/10.1186/s13099-016-0103-7.

Gould, G.M. *et al.* (2018) ‘Detecting clinically actionable variants in the 3’ exons of PMS2 via a reflex workflow based on equivalent hybrid capture of the gene and its pseudogene’, *BMC Medical Genetics*, 19(1), p. 176. Available at: https://doi.org/10.1186/s12881-018-0691-9.

Griffin, D.K. (2022) ‘Why PGT-A, most likely, improves IVF success’, *Reproductive BioMedicine Online*, 45(4), pp. 633–637. Available at: https://doi.org/10.1016/J.RBMO.2022.03.022.

Hafford-Tear, N.J. *et al.* (2019) ‘CRISPR/Cas9-targeted enrichment and long-read sequencing of the Fuchs endothelial corneal dystrophy-associated TCF4 triplet repeat.', *Genetics in medicine : official journal of the American College of Medical Genetics*, 21(9), pp. 2092–2102. Available at: https://doi.org/10.1038/s41436-019-0453-x.

Handyside, A.H. *et al.* (2010) 'Karyomapping: a universal method for genome wide analysis of genetic disease based on mapping crossovers between parental haplotypes', *Journal of Medical Genetics*, 47(10), p. 651. Available at: https://doi.org/10.1136/jmg.2009.069971.

Hardenbol, P. *et al.* (2003) 'Multiplexed genotyping with sequence-tagged molecular inversion probes.', *Nature biotechnology*, 21(6), pp. 673–8. Available at: https://doi.org/10.1038/nbt821.

Hartley, J.L., Zachos, Nicholas C, *et al.* (2010) 'Mutations in TTC37 cause trichohepatoenteric syndrome (phenotypic diarrhea of infancy).', *Gastroenterology*, 138(7), pp. 2388–98, 2398.e1–2. Available at: https://doi.org/10.1053/j.gastro.2010.02.010.

Hartley, J.L., Zachos, Nicholas C., *et al.* (2010) 'Mutations in TTC37 Cause Trichohepatoenteric Syndrome (Phenotypic Diarrhea of Infancy)', *Gastroenterology*, 138(7), pp. 2388-2398.e2. Available at: https://doi.org/10.1053/J.GASTRO.2010.02.010.

Hassan, D. *et al.* (2022) 'Penguin: A tool for predicting pseudouridine sites in direct RNA nanopore sequencing data.', *Methods (San Diego, Calif.)*, 203, pp. 478–487. Available at: https://doi.org/10.1016/j.ymeth.2022.02.005.

Hayes, J.L. *et al.* (2013) 'Diagnosis of copy number variation by Illumina next generation sequencing is comparable in performance to oligonucleotide array comparative genomic hybridisation', *Genomics*. Available at: https://doi.org/10.1016/j.ygeno.2013.04.006.

Hayward, B.E. *et al.* (2007) 'Extensive gene conversion at the PMS2 DNA mismatch repair locus', *Human Mutation*, 28(5). Available at: https://doi.org/10.1002/humu.20457.

Hiatt, J.B. *et al.* (2013) 'Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation.', *Genome research*, 23(5), pp. 843–54. Available at: https://doi.org/10.1101/gr.147686.112.

Hitti-Malin, R.J. *et al.* (2022) ‘Using single molecule Molecular Inversion Probes as a cost-effective, high-throughput sequencing approach to target all genes and loci associated with macular diseases.’, *Human mutation*, 43(12), pp. 2234–2250. Available at: https://doi.org/10.1002/humu.24489.

Höps, W. *et al.* (2025) ‘HiFi long-read genomes for difficult-to-detect, clinically relevant variants.’, *American journal of human genetics*, 112(2), pp. 450–456. Available at: https://doi.org/10.1016/j.ajhg.2024.12.013.

Hosono, S. *et al.* (2003) ‘Unbiased whole-genome amplification directly from clinical samples.’, *Genome research*, 13(5), pp. 954–64. Available at: https://doi.org/10.1101/gr.816903.

Huang, L. *et al.* (2015a) ‘Single-Cell Whole-Genome Amplification and Sequencing: Methodology and Applications.’, *Annual review of genomics and human genetics*, 16, pp. 79–102. Available at: https://doi.org/10.1146/annurev-genom-090413-025352.

Huang, L. *et al.* (2015b) ‘Single-Cell Whole-Genome Amplification and Sequencing: Methodology and Applications’, *Annual Review of Genomics and Human Genetics*, 16. Available at: https://doi.org/10.1146/annurev-genom-090413-025352.

Iannicelli, M. *et al.* (2010) ‘Novel TMEM67 mutations and genotype-phenotype correlates in meckelin-related ciliopathies.’, *Human mutation*, 31(5), pp. E1319-31. Available at: https://doi.org/10.1002/humu.21239.

Idos, G. and Valle, L. (2004) *Lynch Syndrome*.

Illumina (2024) *https://emea.illumina.com/science/genomics-research/articles/constellation-mapped-read-technology.html*.

Illumina (no date) ‘illumina-long-read-prep-data-sheet-m-gl-01420’.

Ingham, D. *et al.* (2013) ‘Simple Detection of Germline Microsatellite Instability for Diagnosis of Constitutional Mismatch Repair Cancer Syndrome’, *Human Mutation*, 34(6), pp. 847–852. Available at: https://doi.org/10.1002/humu.22311.

JACOBS, P.A. and STRONG, J.A. (1959) ‘A case of human intersexuality having a possible XXY sex-determining mechanism.’, *Nature*, 183(4657), pp. 302–3. Available at: https://doi.org/10.1038/183302a0.

Jaganathan, K. *et al.* (2019) ‘Predicting Splicing from Primary Sequence with Deep Learning’, *Cell*, 176(3), pp. 535-548.e24. Available at: https://doi.org/https://doi.org/10.1016/j.cell.2018.12.015.

Jain, M. *et al.* (2018) ‘Nanopore sequencing and assembly of a human genome with ultra-long reads.’, *Nature biotechnology*, 36(4), pp. 338–345. Available at: https://doi.org/10.1038/nbt.4060.

Jia, H. *et al.* (2014) ‘Long-range PCR in next-generation sequencing: comparison of six enzymes and evaluation on the MiSeq sequencer.’, *Scientific reports*, 4, p. 5737. Available at: https://doi.org/10.1038/srep05737.

Jia, H., Tan, S. and Zhang, Y.E. (2024) ‘Chasing Sequencing Perfection: Marching Toward Higher Accuracy and Lower Costs.’, *Genomics, proteomics & bioinformatics*, 22(2). Available at: https://doi.org/10.1093/gpbjnl/qzae024.

Kent, W.J. (2002) ‘BLAT - The BLAST-like alignment tool’, *Genome Research*, 12(4), pp. 656–664. Available at: https://doi.org/10.1101/gr.229202.

Khaddour, R. *et al.* (2007) ‘Spectrum of MKS1 and MKS3 mutations in Meckel syndrome: a genotype-phenotype correlation. Mutation in brief #960. Online.’, *Human mutation*, 28(5), pp. 523–4. Available at: https://doi.org/10.1002/humu.9489.

Kieleczawa, J. (2006) ‘Fundamentals of sequencing of difficult templates--an overview.’, *Journal of biomolecular techniques : JBT*, 17(3), pp. 207–17.

Kokoris, M. *et al.* (2025) ‘Sequencing by Expansion (SBX) – a novel, high-throughput single-molecule sequencing technology’. Available at: https://doi.org/10.1101/2025.02.19.639056.

Lam, C. and Mak, C.M. (2013) ‘Allele dropout caused by a non-primer-site SNV affecting PCR amplification--a call for next-generation primer design algorithm.’, *Clinica chimica acta; international journal of clinical chemistry*, 421, pp. 208–12. Available at: https://doi.org/10.1016/j.cca.2013.03.014.

Lander, E.S. *et al.* (2001) ‘Initial sequencing and analysis of the human genome.’, *Nature*, 409(6822), pp. 860–921. Available at: https://doi.org/10.1038/35057062.

Landrum, M.J. *et al.* (2025) 'ClinVar: updates to support classifications of both germline and somatic variants.', *Nucleic acids research*, 53(D1), pp. D1313–D1321. Available at: https://doi.org/10.1093/nar/gkae1090.

Lasken, R.S. and Stockwell, T.B. (2007) 'Mechanism of chimera formation during the Multiple Displacement Amplification reaction.', *BMC biotechnology*, 7, p. 19. Available at: https://doi.org/10.1186/1472-6750-7-19.

Lee, K.Y. *et al.* (2024) 'Long term outcomes in children with trichohepatoenteric syndrome', *American Journal of Medical Genetics, Part A*, 194(2), pp. 141–149. Available at: https://doi.org/10.1002/ajmg.a.63409.

LEJEUNE, J., GAUTIER, M. and TURPIN, R. (1959) '[Study of somatic chromosomes from 9 mongoloid children].', *Comptes rendus hebdomadaires des seances de l'Academie des sciences*, 248(11), pp. 1721–2.

Leongamornlert, D. *et al.* (2014) 'Frequent germline deleterious mutations in DNA repair genes in familial prostate cancer cases are associated with advanced disease', *British Journal of Cancer*, 110(6), pp. 1663–1672. Available at: https://doi.org/10.1038/bjc.2014.30.

Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics*, 25(14), pp. 1754–1760. Available at: https://doi.org/10.1093/bioinformatics/btp324.

Liao, W.-W. *et al.* (2023) 'A draft human pangenome reference', *Nature*, 617(7960), pp. 312–324. Available at: https://doi.org/10.1038/s41586-023-05896-x.

Lindner, A.K. *et al.* (2021) 'Lynch syndrome: Its impact on urothelial carcinoma', *International Journal of Molecular Sciences*, 22(2). Available at: https://doi.org/10.3390/ijms22020531.

Longmire, J. *et al.* (1987) *A simple salting out procedure for extracting DNA from human nucleated cells*, *Nucleic Acids Research*. C I R L Press Limited.

Lovmar, L. and Syvänen, A.-C. (2006) 'Multiple displacement amplification to create a long-lasting source of DNA for genetic studies', *Human Mutation*, 27(7), pp. 603–614. Available at: https://doi.org/https://doi.org/10.1002/humu.20341.

Macklon, N.S., Geraedts, J.P.M. and Fauser, B.C.J.M. (no date) *Conception to ongoing pregnancy: thèblack box' of early pregnancy loss*.

Madritsch, S. *et al.* (2024) 'Aneuploidy detection in pooled polar bodies using rapid nanopore sequencing.', *Journal of assisted reproduction and genetics*, 41(5), pp. 1261–1271. Available at: https://doi.org/10.1007/s10815-024-03108-7.

Mandelker, D. *et al.* (2016) 'Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing.', *Genetics in medicine : official journal of the American College of Medical Genetics*, 18(12), pp. 1282–1289. Available at: https://doi.org/10.1038/gim.2016.58.

Marcy, Y. *et al.* (2007) 'Nanoliter reactors improve multiple displacement amplification of genomes from single cells', *PLoS Genetics*, 3(9), pp. 1702–1708. Available at: https://doi.org/10.1371/journal.pgen.0030155.

Marosy, B. (2018) *Comparison of Whole Exome Capture Products – Coverage & Quality vs Cost*.

Mc Clinton, B. *et al.* (2023) 'Effective smMIPs-Based Sequencing of Maculopathy-Associated Genes in Stargardt Disease Cases and Allied Maculopathies from the UK.', *Genes*, 14(1). Available at: https://doi.org/10.3390/genes14010191.

McClinton, B., Watson, C.M., *et al.* (2023) 'Haplotyping Using Long-Range PCR and Nanopore Sequencing to Phase Variants: Lessons Learned From the ABCA4 Locus.', *Laboratory investigation; a journal of technical methods and pathology*, 103(8), p. 100160. Available at: https://doi.org/10.1016/j.labinv.2023.100160.

McClinton, B., Crinnion, L.A., *et al.* (2023) 'Targeted nanopore sequencing enables complete characterisation of structural deletions initially identified using exon-based short-read sequencing strategies.', *Molecular genetics & genomic medicine*, 11(6), p. e2164. Available at: https://doi.org/10.1002/mgg3.2164.

Mullis, K. *et al.* (1986) 'Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction.', *Cold Spring Harbor symposia on quantitative biology*, 51 Pt 1, pp. 263–73. Available at: https://doi.org/10.1101/sqb.1986.051.01.032.

Muraro, M.J. *et al.* (2016) 'A Single-Cell Transcriptome Atlas of the Human Pancreas.', *Cell systems*, 3(4), pp. 385-394.e3. Available at: https://doi.org/10.1016/j.cels.2016.09.002.

Musto, H. *et al.* (1997) 'Compositional constraints in the extremely GC-poor genome of Plasmodium falciparum.', *Memorias do Instituto Oswaldo Cruz*, 92(6), pp. 835–41. Available at: https://doi.org/10.1590/s0074-02761997000600020.

Ndugga-Kabuye, M.K. and Issaka, R.B. (2019) 'Inequities in multi-gene hereditary cancer testing: lower diagnostic yield and higher VUS rate in individuals who identify as Hispanic, African or Asian and Pacific Islander as compared to European.', *Familial cancer*, 18(4), pp. 465–469. Available at: https://doi.org/10.1007/s10689-019-00144-6.

Olave, M.C. and Graham, R.P. (2022) 'Mismatch repair deficiency: The what, how and why it is important', *Genes, Chromosomes and Cancer*, 61(6), pp. 314–321. Available at: https://doi.org/https://doi.org/10.1002/gcc.23015.

Ospino, M.C. *et al.* (2024) 'Evaluation of multiple displacement amplification for metagenomic analysis of low biomass samples.', *ISME communications*, 4(1), p. ycae024. Available at: https://doi.org/10.1093/ismeco/ycae024.

Oxford Nanopore Technologies (2024) *https://github.com/nanoporetech/dorado?tab=readme-ov-file#dna-models*.

Oxford Nanopore Technologies (no date a) *https://nanoporetech.com/document/requirements/varying-pcr-cycles*.

Oxford Nanopore Technologies (no date b) *Sequencing products of multiple displacement amplification (MDA)*.

PabBio (2025) *https://www.pacb.com/wp-content/uploads/Application-note-Comprehensive-genotyping-with-the-PureTarget-repeat-expansion-panel-and-HiFi-sequencing.pdf*.

PacBio (2025) *https://www.pacb.com/wp-content/uploads/Revio-brochure.pdf*.

Pan, S. *et al.* (no date) *Common variants in PMS2CL that can present in PMS2 as pathogenic variants with extremely low frequencies*.

Pascual, J. *et al.* (2022) 'ESMO recommendations on the use of circulating tumour DNA assays for patients with cancer: a report from the ESMO Precision Medicine Working Group.', *Annals of oncology : official journal of the European Society for Medical Oncology*, 33(8), pp. 750–768. Available at: https://doi.org/10.1016/j.annonc.2022.05.520.

Pengelly, R.J. *et al.* (2013) ‘A SNP profiling panel for sample tracking in whole-exome sequencing studies.’, *Genome medicine*, 5(9), p. 89. Available at: https://doi.org/10.1186/gm492.

Perez, G. *et al.* (2025) ‘The UCSC Genome Browser database: 2025 update.’, *Nucleic acids research*, 53(D1), pp. D1243–D1249. Available at: https://doi.org/10.1093/nar/gkae974.

Pinard, R. *et al.* (2006) ‘Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing.’, *BMC genomics*, 7, p. 216. Available at: https://doi.org/10.1186/1471-2164-7-216.

Porreca, G.J. *et al.* (2007) ‘Multiplex amplification of large sets of human exons.’, *Nature methods*, 4(11), pp. 931–6. Available at: https://doi.org/10.1038/nmeth1110.

Rebuzzi, F., Ulivi, P. and Tedaldi, G. (2023) ‘Genetic Predisposition to Colorectal Cancer: How Many and Which Genes to Test?’, *International Journal of Molecular Sciences*. MDPI. Available at: https://doi.org/10.3390/ijms24032137.

Rhee, M. *et al.* (2016) ‘Digital Droplet Multiple Displacement Amplification (ddMDA) for Whole Genome Sequencing of Limited DNA Samples.’, *PloS one*, 11(5), p. e0153699. Available at: https://doi.org/10.1371/journal.pone.0153699.

Richards, S. *et al.* (2015) ‘Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.’, *Genetics in medicine : official journal of the American College of Medical Genetics*, 17(5), pp. 405–24. Available at: https://doi.org/10.1038/gim.2015.30.

Robinson, J.T. *et al.* (2011) ‘Integrative genomics viewer’, *Nature Biotechnology*, 29(1), pp. 24–26. Available at: https://doi.org/10.1038/nbt.1754.

Sample to Insight EZ1&2 DNA Tissue Handbook (2022).

Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *DNA sequencing with chain-terminating inhibitors (DNA polymerase/nucleotide sequences/bacteriophage 4X174)*.

Segura, A.V.C. *et al.* (2024) ‘Misclassification of a frequent variant from PMS2CL pseudogene as a PMS2 loss of function variant in Brazilian patients.’, *Familial* *cancer*, 23(4), pp. 653–657. Available at: https://doi.org/10.1007/s10689-024-00411-1.

Sidore, A.M. *et al.* (2016) ‘Enhanced sequencing coverage with digital droplet multiple displacement amplification.’, *Nucleic acids research*, 44(7), p. e66. Available at: https://doi.org/10.1093/nar/gkv1493.

Sigalos, G.A., Triantafyllidou, O. and Vlahos, N.F. (2016) ‘Novel embryo selection techniques to increase embryo implantation in IVF attempts’, *Archives of Gynecology and Obstetrics*, 294(6), pp. 1117–1124. Available at: https://doi.org/10.1007/s00404-016-4196-5.

da Silva Franco, J.F. *et al.* (2025) ‘Hidden Aberrant Transcripts in TTC37 Cause Trichohepatoenteric Syndrome.’, *Clinical genetics*, 107(1), pp. 113–114. Available at: https://doi.org/10.1111/cge.14630.

Simpson, J.L., Kuliev, A. and Rechitsky, S. (2019) ‘Overview of Preimplantation Genetic Diagnosis (PGD): Historical Perspective and Future Direction’, in B. Levy (ed.) *Prenatal Diagnosis*. New York, NY: Springer New York, pp. 23–43. Available at: https://doi.org/10.1007/978-1-4939-8889-1_2.

Smith, U.M. *et al.* (2006) ‘The transmembrane protein meckelin (MKS3) is mutated in Meckel-Gruber syndrome and the wpk rat.’, *Nature genetics*, 38(2), pp. 191–6. Available at: https://doi.org/10.1038/ng1713.

de Sousa Dias, M. *et al.* (2013) ‘Detection of novel mutations that cause autosomal dominant retinitis pigmentosa in candidate genes by long-range PCR amplification and next-generation sequencing.’, *Molecular vision*, 19, pp. 654–64.

Stark, Z. *et al.* (2021) ‘Scaling national and international improvement in virtual gene panel curation via a collaborative approach to discordance resolution’, *The American Journal of Human Genetics*, 108(9), pp. 1551–1557. Available at: https://doi.org/https://doi.org/10.1016/j.ajhg.2021.06.020.

Steyaert, W. *et al.* (2024) ‘Unravelling undiagnosed rare disease cases by HiFi long-read genome sequencing.’, *medRxiv : the preprint server for health sciences* [Preprint]. Available at: https://doi.org/10.1101/2024.05.03.24305331.

Stjepanovic, N. *et al.* (2019) ‘Hereditary gastrointestinal cancers: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up†.’, *Annals of oncology :*

official journal of the European Society for Medical Oncology, 30(10), pp. 1558–1571. Available at: https://doi.org/10.1093/annonc/mdz233.

Sung, H. *et al.* (2021) 'Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries.', *CA: a cancer journal for clinicians*, 71(3), pp. 209–249. Available at: https://doi.org/10.3322/caac.21660.

Tecan (no date) *WALK-AWAY SOLUTION THAT DELIVERS UP TO 96 CONSISTENT DNA LIBRARIES COMPATIBLE WITH ALL OXFORD NANOPORE SEQUENCING DEVICES preparation with DreamPrep ® NGS and Ligation Sequencing Kit XL V14 for nanopore sequencing High-throughput DNA library*.

Telenius, H. *et al.* (1992) 'Degenerate oligonucleotide-primed PCR: General amplification of target DNA by a single degenerate primer', *Genomics*, 13(3), pp. 718–725. Available at: https://doi.org/https://doi.org/10.1016/0888-7543(92)90147-K.

TJIO, J.O.E.H.I.N. and LEVAN, A. (1956) 'THE CHROMOSOME NUMBER OF MAN', *Hereditas*, 42(1–2), pp. 1–6. Available at: https://doi.org/https://doi.org/10.1111/j.1601-5223.1956.tb03010.x.

Twist Bioscience (no date) *Long-Read Library Preparation and Standard Hyb v2 Enrichment*.

Vermeulen, C. *et al.* (2023) 'Ultra-fast deep-learned CNS tumour classification during surgery', *Nature*, 622(7984), pp. 842–849. Available at: https://doi.org/10.1038/s41586-023-06615-2.

De Vos, M. *et al.* (2004) 'Novel PMS2 Pseudogenes Can Conceal Recessive Mutations Causing a Distinctive Childhood Cancer Syndrome', *The American Journal of Human Genetics*, 74(5), pp. 954–964. Available at: https://doi.org/https://doi.org/10.1086/420796.

Walker, L.C. *et al.* (2023) 'Using the ACMG/AMP framework to capture evidence related to predicted and observed impact on splicing: Recommendations from the ClinGen SVI Splicing Subgroup', *The American Journal of Human Genetics*, 110(7), pp. 1046–1067. Available at: https://doi.org/https://doi.org/10.1016/j.ajhg.2023.06.002.

Watson, C.M. *et al.* (2014) ‘Diagnostic whole genome sequencing and split-read mapping for nucleotide resolution breakpoint identification in CNTNAP2 deficiency syndrome.’, *American journal of medical genetics. Part A*, 164A(10), pp. 2649–55. Available at: https://doi.org/10.1002/ajmg.a.36679.

Watson, C.M. *et al.* (2016a) ‘Enhanced diagnostic yield in Meckel-Gruber and Joubert syndrome through exome sequencing supplemented with split-read mapping.’, *BMC medical genetics*, 17, p. 1. Available at: https://doi.org/10.1186/s12881-015-0265-z.

Watson, C.M. *et al.* (2016b) ‘Enhanced diagnostic yield in Meckel-Gruber and Joubert syndrome through exome sequencing supplemented with split-read mapping.’, *BMC medical genetics*, 17, p. 1. Available at: https://doi.org/10.1186/s12881-015-0265-z.

Watson, C.M. *et al.* (2017) ‘Characterization and Genomic Localization of a SMAD4 Processed Pseudogene.’, *The Journal of molecular diagnostics : JMD*, 19(6), pp. 933–940. Available at: https://doi.org/10.1016/j.jmoldx.2017.08.002.

Watson, C.M., Crinnion, L.A., *et al.* (2020a) ‘Cas9-based enrichment and single-molecule sequencing for precise characterization of genomic duplications.’, *Laboratory investigation; a journal of technical methods and pathology*, 100(1), pp. 135–146. Available at: https://doi.org/10.1038/s41374-019-0283-0.

Watson, C.M., Crinnion, L.A., *et al.* (2020b) ‘Cas9-based enrichment and single-molecule sequencing for precise characterization of genomic duplications.’, *Laboratory investigation; a journal of technical methods and pathology*, 100(1), pp. 135–146. Available at: https://doi.org/10.1038/s41374-019-0283-0.

Watson, C.M., Dean, P., *et al.* (2020) ‘Long-read nanopore sequencing resolves a TMEM231 gene conversion event causing Meckel-Gruber syndrome.’, *Human mutation*, 41(2), pp. 525–531. Available at: https://doi.org/10.1002/humu.23940.

Watson, C.M. *et al.* (2022) ‘Long-read sequencing to resolve the parent of origin of a de novo pathogenic UBE3A variant.’, *Journal of medical genetics*, 59(11), pp. 1082–1086. Available at: https://doi.org/10.1136/jmedgenet-2021-108314.

Wimmer, K. *et al.* (2014) ‘Diagnostic criteria for constitutional mismatch repair deficiency syndrome: suggestions of the European consortium “care for CMMRD”

(C4CMMRD).', *Journal of medical genetics*, 51(6), pp. 355–65. Available at: https://doi.org/10.1136/jmedgenet-2014-102284.

Woyke, T. *et al.* (2011) 'Decontamination of MDA reagents for single cell whole genome amplification.', *PloS one*, 6(10), p. e26161. Available at: https://doi.org/10.1371/journal.pone.0026161.

Yao, N.Y. and O'Donnell, M. (2012) 'The RFC clamp loader: Structure and Function', *Subcellular Biochemistry*, 62, pp. 259–279. Available at: https://doi.org/10.1007/978-94-007-4572-8_14.

Yu, J.-H. *et al.* (2023) 'Efficacy of PD-1 inhibitors for colorectal cancer and polyps in Lynch syndrome patients.', *European journal of cancer (Oxford, England : 1990)*, 192, p. 113253. Available at: https://doi.org/10.1016/j.ejca.2023.113253.

Zeng, T. and Li, Y.I. (2022) 'Predicting RNA splicing from DNA sequence using Pangolin', *Genome Biology*, 23(1), p. 103. Available at: https://doi.org/10.1186/s13059-022-02664-4.

Zhang, X. *et al.* (2021) 'Annotating high-impact 5'untranslated region variants with the UTRannotator.', *Bioinformatics (Oxford, England)*, 37(8), pp. 1171–1173. Available at: https://doi.org/10.1093/bioinformatics/btaa783.

Zheng, G.X.Y. *et al.* (2016) 'Haplotyping germline and cancer genomes with high-throughput linked-read sequencing', *Nature Biotechnology*, 34(3), pp. 303–311. Available at: https://doi.org/10.1038/nbt.3432.

Zhou, B. *et al.* (2024) 'Resolving the 22q11.2 deletion using CTLR-Seq reveals chromosomal rearrangement mechanisms and individual variance in breakpoints.', *Proceedings of the National Academy of Sciences of the United States of America*, 121(31), p. e2322834121. Available at: https://doi.org/10.1073/pnas.2322834121.

Zhou, G. *et al.* (2022) 'Performance characterization of PCR-free whole genome sequencing for clinical diagnosis.', *Medicine*, 101(10), p. e28972. Available at: https://doi.org/10.1097/MD.0000000000028972.

Zong, C. *et al.* (2012) 'Genome-Wide Detection of Single-Nucleotide and Copy-Number Variations of a Single Human Cell', *Science*, 338(6114), pp. 1622–1626. Available at: https://doi.org/10.1126/science.1229164.

Zook, J.M. *et al.* (2019) ‘An open resource for accurately benchmarking small variant and reference calls’, *Nature Biotechnology*, 37(5), pp. 561–566. Available at: https://doi.org/10.1038/s41587-019-0074-6.