UNIVERSITA' DEGLI STUDI DI VERONA

DEPARTMENT OF

Biotechnology

GRADUATE SCHOOL OF

Natural Sciences and Engineering

DOCTORAL PROGRAM IN

Biotechnology

XXXIV cycle

TITLE OF THE DOCTORAL THESIS

Evaluation and optimization of long-DNA capture approaches for the characterization of long microsatellites in repeat expansion disorders

S.S.D. BIO/18

Coordinator: Prof. Matteo Ballottari

Tutor: Prof. Marzia Rossato

Doctoral Student: Massimiliano Alfano

Evaluation and optimization of long-DNA capture approaches for the characterization of long microsatellites in repeat expansion disorders
– Massimiliano Alfano
PhD thesis
Verona, July the 26th 2022
ISBN xxx

SUMMARY

L'avvento del sequenziamento a read lunghe ha migliorato la nostra capacità di caratterizzare regioni genomiche complesse caratterizzate da grandi variazi strutturali, elementi ripetuti, contenuto anormale di GC o geni altamente omologhi. La combinazione del sequenziamento a read lunghe con strategie di arricchimento che consentono di catturare frammenti lunghi di DNA rappresenta uno strumento prezioso per ridurre i costi di analisi, massimizzando la produzione di dati su una regione di interesse specifica. In questa tesi sono state valutate le caratteristiche e le prestazioni di tre diverse metodologie di cattura di frammenti di DNA lungo, tra i quali figurano la cattura indiretta (Xdrop di Samplix), il sequenziamento target mediato da Cas9 ed una serie di metodi basati su ibridazione mediante sonde (PNA, dCas9 e dsDNA-probes). I vantaggi di questi approcci sono stati valutati in combinazione con il sequenziamento a read lunghe mediante nanopori per l'analisi delle ripetizioni presenti nei geni *FMR1*, *DMPK* e *CNBP*. Gli stessi sono stati selezionati come casi di studio di malattie congenite contraddistinte rispettivamente da target medio-lunghi, lunghi e ultra-lunghi.

Tutti i metodi hanno permesso efficacemente l'arricchimento di frammenti lunghi di DNA oltre le kilobasi, anche se loro lunghezza è risultata variabile tra i diversi approcci. In particolare, la cattura mediata da Cas9 ha permesso di sequenziare molecole fino a 50 kbp di lunghezza, grazie alle quali è stato possibile caratterizzare una ripetizione di 46.6 kbp, una delle più lunghe ottenute con approcci di arricchimento. Nonostante si sia rivelato l'approccio più sensibile alla qualità del DNA di partenza, la cattura mediata da Cas9 ha consentito di ottenere il più alto arricchimento tra i metodi testati. Usando il metodo Xdrop siamo riusciti a catturare porzioni di DNA ancora più lunghe (100 kbp) rispetto a Cas9, seppur frammentate in porzioni da circa 5-10 kbp, e ad un livello di arricchimento leggermente inferiore. Inoltre, il metodo Xdrop ha permesso di lavorare con le quantità più basse di DNA di partenza (10 ng), rispetto ai microgrammi di DNA necessari per gli altri metodi, suggerendo pertanto il potenziale di questo approccio per campioni derivati da test pre-impianto o pre-natali, biopsie cliniche o anche singole cellule. Uno svantaggio legato all'uso di basse quantità di DNA di partenza è stata la necessità di amplificare con metodo WGA (whole genome amplification) il DNA arricchito. Questa procedura ha infatti diminuito le dimensioni

delle molecole sequenziate, limitando quindi l'applicabilità di questo metodo. Inoltre, l'approccio Xdrop richiede gli investimenti iniziali più elevati, in quanto prevede l'utilizzo di un generatore di droplet e di un citometro a flusso per l'arricchimento. Gli approcci basati sull'ibridazione potrebbero rappresentare la soluzione più conveniente, con costi di circa 10 volte inferiori rispetto a Xdrop e Cas9. Tuttavia, utilizzando questi metodi il recupero del DNA arricchito è stato così basso da non consentire la successiva analisi di sequenziamento con read lunghe. Anche se gli approcci basati sull'ibridazione rappresentano soluzioni potenzialmente interessanti, la loro efficace applicazione richiederà successive ottimizzazioni volte a migliorare la resa del DNA arricchito.

Gli approcci Xdrop e Cas9 hanno consentito il sequenziamento a read lunghe, usando la piattaforma Oxford Nanopore Technology, di DNA estratto da pazienti che presentavano espansioni patogeniche. Ciò ha dimostrato la capacità di questi approcci di catturare e caratterizzare microsatelliti significativamente lunghi, ed in modo consistente rispetto agli approcci diagnostici tradizionali. Inoltre, i metodi hanno consentito l'accurata discriminazione di alleli normali ed espansi, con la valutazione simultanea della lunghezza, struttura e motivo della ripetizione e del livello di mosaicismo somatico. Informazioni non ottenibili simultaneamente non con i metodi tradizionali (usati da soli o in combinazione). L'applicazione di questi approcci di cattura di DNA lungo in ambito clinico potrebbe potenzialmente migliorare la diagnosi dei pazienti e fornire correlazioni genotipo-fenotipo più precise, ancora carenti/limitate per quelle malattie caratterizzate da grandi espansioni di microsatelliti.

In conclusione, la valutazione approfondita dei punti di forza e di debolezza degli approcci di cattura del DNA lungo – come descritto in questa tesi – promuoverà la loro applicazione più diffusa per la caratterizzazione di loci patogenici ancora solo parzialmente esplorati utilizzando gli approcci tradizionali.

ABSTRACT

The advent of long-read sequencing has enhanced our capability to characterize complex genomic regions harboring large structural variations, repetitive elements, abnormal GC content or highly homologous genes. The combination of long-read sequencing with enrichment strategies that allow to capture long fragments represents a valuable tool to reduce analysis costs while maximizing data production on a selected region of interest. Here we evaluated the features and performances of three different long-DNA capture approaches, comprising indirect sequence capture (Samplix's Xdrop), Cas9-mediated targeted sequencing, and a set of three hybridization-capture methods (PNA, dCas9 and dsDNA-probes). The benefits of these approaches in combination with long-read sequencing were assessed for the analysis of *FMR1*, *DMPK*, and *CNBP* repeat expansions, selected as case studies for medium-long, long and ultra-long targets, respectively, and causative of congenital disorders.

All methods resulted in successful enrichment of long DNA target molecules, even if the length of enriched DNA was variable across the different approaches. In particular the Cas9-mediated capture allowed to sequence up to 50 kbp molecules in length, and thus to characterize a repeat of 46.6 kbp, one of the longest achieved with target-enrichment approaches. Despite being the most sensitive approach to input gDNA quality, Cas9-mediated capture enabled also to achieve the highest fold-enrichment among the three methods tested. Using the Xdrop-mediated method, we could capture even longer DNA portions (100 kbp) than with Cas9, at just slightly lower enrichment level, even if these were fragmented in shorter pieces of 5-10 kbp in length. In addition, the Xdrop method allowed to work with the lowest gDNA input (10 ng), in contrast to the micrograms of gDNA required for the other methods, suggesting the potential of this approach to work with samples derived from pre-natal/pre-implant testing, clinical biopsies or even single cells. A drawback linked to the use of lower input was the need of Whole Genome Amplification downstream the enrichment step, that decreased the size of sequenced molecules and thus restrained the applicability of this method. Also, the Xdrop approach required the highest capital investments, as it depends on a specific droplet generator instrument and a cytometer for sorting. The hybridization-based approaches could potentially represent the most cost-effective solution, with costs ~10 times lower than Xdrop and Cas9. However, using these

methods the downstream recovery of enriched DNA was so low that did not allow the subsequent sequencing analysis with long-reads. Even if hybridization-based approaches represent potentially interesting solutions, subsequent protocol optimization aimed at improving the yield of enriched-DNA are therefore required for their effective application.

Xdrop- and Cas9-mediated workflows enabled successful ONT sequencing of patient genomic DNA harboring known repeat expansions. This demonstrated the capability of these approaches to capture and characterize significantly long (pathogenic) microsatellites, in high agreement with traditional diagnostic approaches. In addition, the methods allowed to achieve the accurate discrimination at single nucleotide resolution of normal and expanded alleles, with the simultaneous assessment of repeat length, structure/motif and level of somatic mosaicism, otherwise not feasible with traditional methods (used either alone or in combination). Application of these long-DNA capture approaches in the clinical setting could potentially improve patient diagnosis and provide more precise genotype-phenotype correlations, which are still lacking/limited for those disease characterized by large microsatellite expansions.

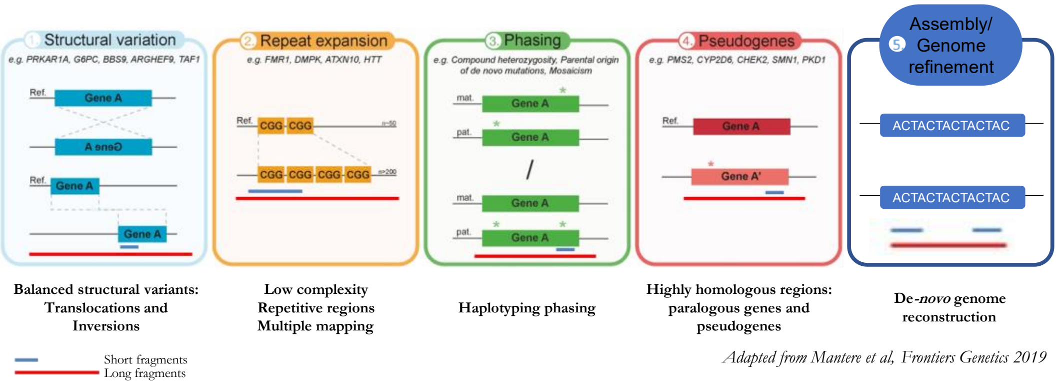
In conclusion, the deep evaluation of strengths and weaknesses of long-DNA capture approaches – as described in this thesis – will promote their more widespread application for the characterization of pathogenic loci, only partially resolved using traditional approaches.

TABLE OF CONTENTS

1. INTRODUCTION

1.1 THE BENEFITS OF ENRICHING LONG DNA FRAGMENTS

The advent of next-generation sequencing (NGS) has revolutionized our understanding of the genome, mostly due to a consistent increase in throughput and accuracy, associated to a significant reduction of the costs. This enabled a more comprehensive and accurate characterization of various genomic regions, reason why NGS platforms have been widely used for disease characterization. State of the art NGS platforms usually rely on short-reads and PCR-based workflows. However, the limited read length (< 600 bp) and lack of contextual information has restricted their utility mainly to the characterization of short nucleotide variations (SNV), short insertions and deletions (indels). However, as recently pointed out by Ebbert et al.[1], the human genome contains several "dark" regions that are invisible to short-read sequencing, whose analysis is important for several clinical conditions (**Figure 1**).


1. Structural variation
e.g. PRKAR1A, G6PC, BBS9, ARGHEF9, TAF1
Ref.
Gene A
A eneG
Ref.
Gene A
Gene A
2. Repeat expansion
e.g. FMR1, DMPK, ATXN10, HTT
Ref.
CGG CGG
n~50
CGG CGG CGG CGG
n>200
3. Phasing
e.g. Compound heterozygosity, Parental origin of de novo mutations, Mosaicism
mat.
Gene A
pat.
Gene A
/
mat.
Gene A
pat.
Gene A
4. Pseudogenes
e.g. PMS2, CYP2D6, CHEK2, SMN1, PKD1
Ref.
Gene A
Gene A'
Assembly/ 5. Genome refinement
ACTACTACTACTAC
ACTACTACTACTAC
Balanced structural variants: Translocations and Inversions
Low complexity Repetitive regions Multiple mapping
Haplotyping phasing
Highly homologous regions: paralogous genes and pseudogenes
De-*novo* genome reconstruction
Short fragments
Long fragments

Adapted from Mantere et al, Frontiers Genetics 2019

Figure 1. Overview of genomic dark regions and the benefits of long fragments for their characterization. (1) Improved characterization of balanced structural variants. **(2)** Capability to span across entire repeat expansions, including the flanking sequences for repeat size precise assessment. **(3)** Enhanced haplotype phasing, i.e., assignment of genetic variants to paternal or maternal chromosomes. **(4)** Improved discrimination of clinically relevant genes from their pseudogenes. **(5)** Improved *de-novo* genome assembly and genome refinement (e.g. gap-filling).

Despite the use of sophisticated bioinformatic algorithms, it is often impossible to accurately map, or even assemble, short reads originating from regions harboring large structural variations (SVs), repetitive sequences, abnormal (>60%) GC content or highly homologous genes (e.g. pseudogenes)[2,3]. Also, haplotype phasing has been proven difficult when variants' distance exceeds maximal read length[4]. As a result, technologies that provide a substantial increase in the read length are highly desirable to resolve structurally complex regions. At the cost of a lower accuracy, long-reads provide contextual information as they span the entire complex region, including its flanking sequences.

The most widely used long-read sequencing platforms are provided by Pacific Biosciences (PacBio, Menlo Park, California, USA) and Oxford Nanopore technology (ONT, Oxford, UK). PacBio sequencing can generate either continuous long reads (CLRs) or high fidelity (HiFi) reads. CLRs are typically 20-60 kbp in length[5,6] with error rates ranging from 8% to 15%[5,7]. HiFi reads are generated by collapsing multiple reads originating from the same template, thus compensating errors, that decrease down to 1% or less[5,7,8], but at the cost of a reduced read length (10 – 25 kbp[8]).

On the other hand, ONT sequencing routinely generates reads from 10 – 30 kbp up to megabases, with the current record held at 2.3 Mb[9]. The current error rate is lower as compared to PacBio (< 5%[10]) sequencing, but less randomic and more frequent in homopolymers. Also, ONT has recently released a new sequencing chemistry (Q20+) which, in combination with the latest nanopore version (R10.4), has shown a consistent reduction of the error rate down to 1% - 1.7%[11,12].

Still, long-read sequencing technologies are relatively new, reason why they still suffer from high costs and are still confined to the research field. Sequencing a whole human genome (3.2 Gb) at 30x coverage using the Illumina platform (short reads) costs € 900, while the equivalent analysis with PacBio and Nanopore cost € 5,500 and € 3,500, respectively. Hence, the combination of long-read sequencing technologies with enrichment strategies that allow to capture long fragments is highly desirable, to reduce costs while ensuring sufficient coverage on the target site for accurate characterization. Targeted sequencing also allows faster and simpler downstream bioinformatics analysis which are more compatible with diagnostic applications.

1.2 LONG-DNA FRAGMENT ENRICHMENT APPROACHES

Several long-DNA-fragment enrichment approaches are available, their basic principle being sometimes very different (**Table 1**). However, their application is limited to the research setting and an extensive benchmarking of their performances for the characterization of difficult genomic regions is still absent.

Table 1. Overview of long-DNA-fragment enrichment approaches

Method	Enrichment principle
Peptide nucleic acids	PNA hybridization and pull-down
dCas9	Binding specificity of nuclease-deficient Cas9 and pull-down
DNA-probes	DNA hybridization and pull-down
Region-specific extraction	Enzymatic extension and pull-down
Type IIS restriction enzymes	Sequence specific overhangs
Xdrop	Digital encapsulation and flow sorting
Chromosome sorting	Flow sorting
Cas9	Cas9-driven cut and sequencing adapter ligation at specific target sites
CATCH	Cas9-driven target cut and isolation of DNA fragments by Pulsed field gel electrophoresis

1.2.1 Hybridization-based capture

Hybridization capture takes advantage of the hybridization of DNA/RNA probes to a region of interest followed by target pull-down using mostly streptavidin beads. In general, these methods have been optimized mostly for clinical applications and for the capture of short fragments. Due to the low recovery of target DNA, they have often been coupled either to canonical PCR[13–18] or whole genome amplification (WGA)[19], followed by Illumina short-read sequencing. Some approaches, however, have also been employed to pull-down long DNA fragments of 10 – 60 kbp[19–21] and involved the use of Peptide nucleic acids (PNA) and nuclease-deficient "dead" Cas9 (dCas9).

Peptide nucleic acids. PNAs (**Figure 2A**) were first introduced by Nielsen et al.[22] and consist of short (12 – 21 bases) sequences harboring an uncharged pseudo-peptide polymer backbone which confers higher affinity to DNA due to the lack of repulsive forces. Such structure also confers higher melting temperature (Tm) and stability as compared to DNA-probes[22,23,20]. Previous reports have shown the use of a single biotinylated PNA for the PCR-free capture of genomic regions up to 10 - 60 kbp in size[20,21] coupled to Sanger sequencing. Importantly, no reports have been produced to date which coupled PCR-free PNA-based capture to long-read sequencing.

Dead-Cas9. dCas9 (**Figure 2B**) is a recombinant nuclease that carries knock-out mutations in the RuvC-like and HNH nuclease domains[24]. When assembled to a gRNA, the "dead" ribonucleoprotein (dRNP) complex retains the ability to bind to the target without cleaving it. The system has been largely exploited for genomic visualization via the fusion with fluorescent proteins, gene regulation through fusion with activators or repressors, alteration in epigenetic modifications through fusion with methyltransferases or deacetylases, and immunoprecipitation[24–26]. In a recent report, biotinylated dCas9/gRNAs were used to pull down gDNA fragment of up to 15 kbp in length[19], followed by WGA and Illumina short-read sequencing. To date, no report has been produced which employed biotinylated dCas9 for the capture of long-DNA fragments coupled to long-read sequencing.

DNA-probes. DNA probes (**Figure 2C**) are short (~20 bp) probes tiling across a given target region. Methods involving biotinylated DNA/RNA-probes have been extensively compared and optimized to capture short-DNA fragments for human exome characterization and using the Illumina platform[13–18]. A recent report also described the use of DNA probes for the capture of long-DNA fragments coupled to ONT sequencing, enabling the enrichment of complete plastid genomes[27].

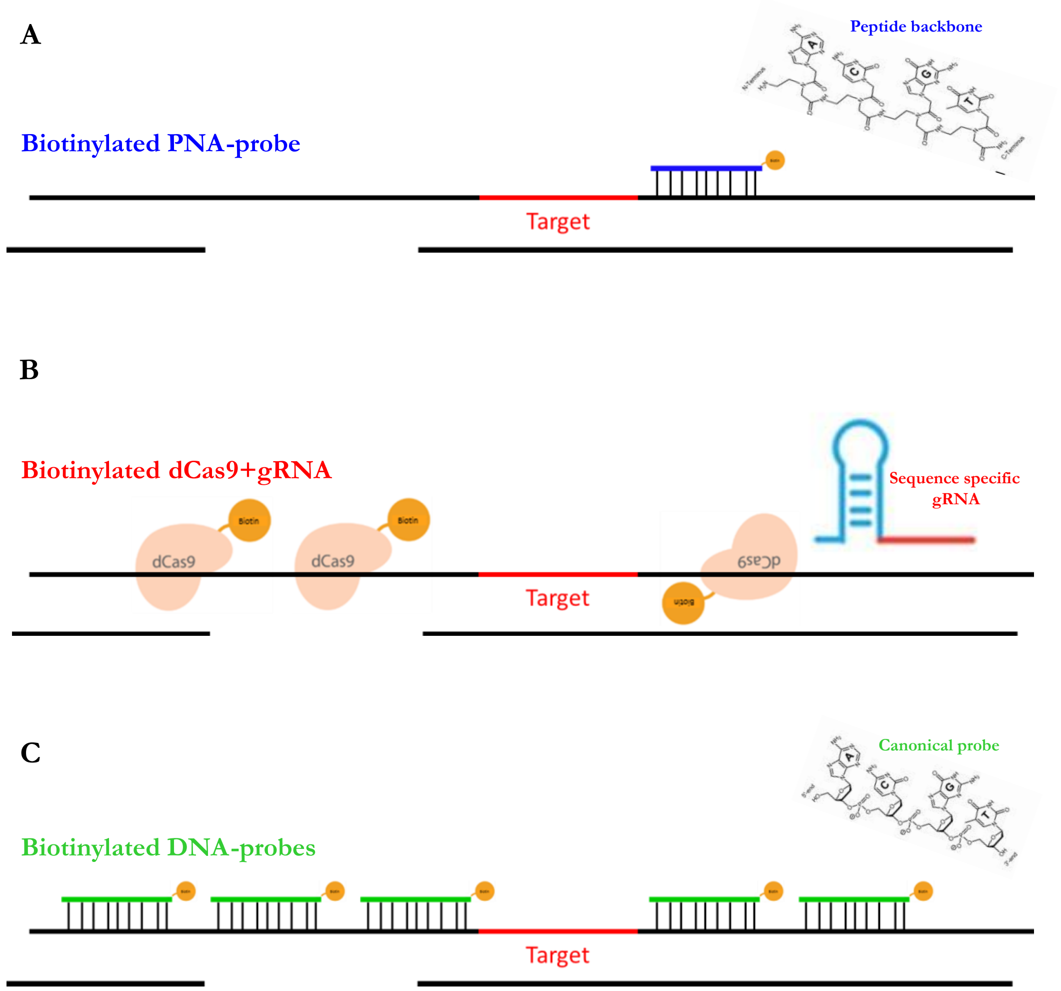


Figure 2. Schematic representation of hybridization-based approaches. (A) Biotinylated peptide nucleic acid probes (PNAs). The presence of an uncharged pseudo-peptide polymer backbone confers higher affinity to DNA due to the lack of repulsive forces as well as higher melting temperature and stability. Due to the higher affinity and stability, only one PNA is potentially necessary for target pull-down. **(B)** Biotinylated nuclease-deficient “dead” Cas9 (dCas9). When assembled to a gRNA, the dRNP complex retains the ability to bind to the target without cleaving it. The biotin on the recombinant dCas9 or synthetic gRNA is then used for target pull-down. **(C)** Biotinylated DNA-probes tiling across the target region. For all methods, target hybridization was followed by Streptavidin-mediated pull-down.

1.2.2 Region-specific extraction

Region-specific extraction (**Figure 3**) relies on the hybridization of short (20-25 bp) oligonucleotide primers to a given target region. These primers are then enzymatically extended, incorporating biotinylated nucleotides into the newly synthetized DNA. The original target molecule is then pulled-down by using streptavidin-coated beads. The method has been first described by Dapprich et al.[28] and was coupled to WGA for the capture of overlapping 20 kbp DNA chunks which spanned a 4 Mbp portion of the major histocompatibility complex. Illumina sequencing of the captured DNA generated up to 164x coverage on the entire target (MHC). The WGA step allows to start from limited amount of input DNA (< 500 ng), and since the yield can be consistently high (10 - 40 μg), the method has also the potential to be coupled to long-read sequencing platforms.

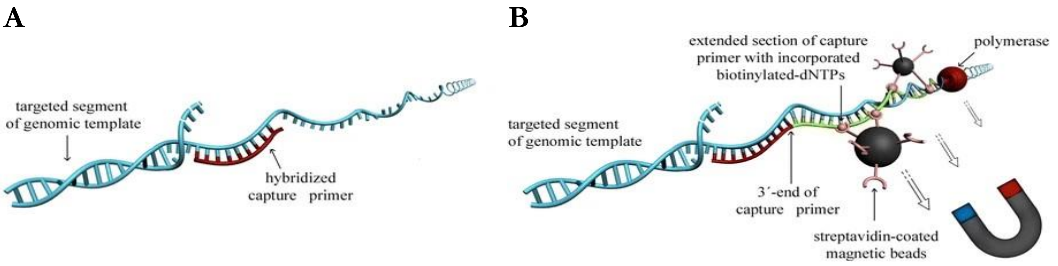

A
targeted segment of genomic template
hybridized capture primer
B
extended section of capture primer with incorporated biotinylated-dNTPs
polymerase
targeted segment of genomic template
3´-end of capture primer
streptavidin-coated magnetic beads

Figure 3. Principle of region-specific extraction. (A) Capture primers are hybridized to the target region. **(B)** Enzymatic extensions using biotinylated dNTPs, followed by streptavidin-mediated pull-down. *Retrieved from Dapprich et al, BMC Genomics 2016*

1.2.3 Type IIS restriction enzymes

The method exploits type-IIS restriction endonucleases (REs, **Figure 4**), which are known to cut at a specific distance outside of their recognition sequence, usually within 1-20 nucleotides[29,30]. This produces DNA fragments with a single-stranded overhang sequence determined only by the local context at the site of cleavage. Two independent hairpin adapters are ligated to the target's ends. Adapter ligation originates closed circular DNA, which is refractory to subsequent digestion with exonucleases. Off-target circular DNA is instead removed by using additional REs, not cutting the region of interest. The method has been applied

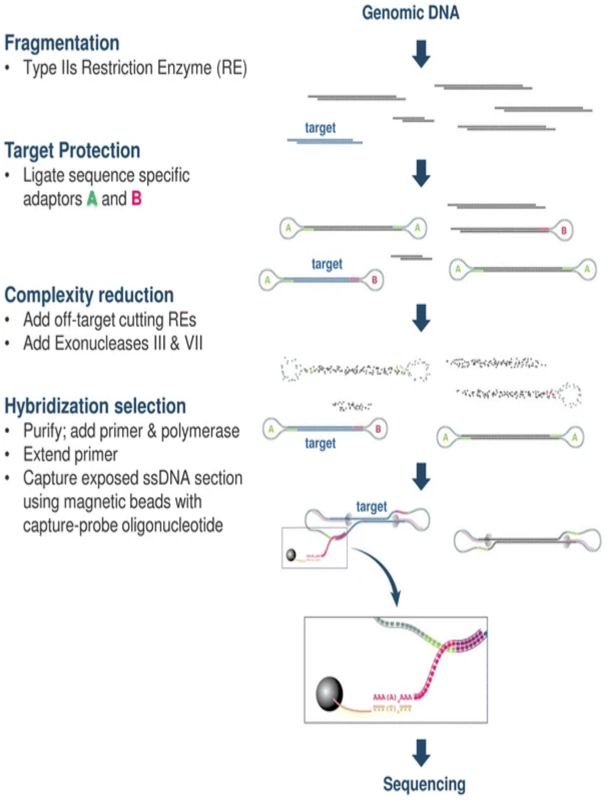

Genomic DNA
Fragmentation
• Type IIs Restriction Enzyme (RE)
target
Target Protection
• Ligate sequence specific adaptors A and B
A
B
target
Complexity reduction
• Add off-target cutting REs
• Add Exonucleases III & VII
Hybridization selection
• Purify; add primer & polymerase
• Extend primer
• Capture exposed ssDNA section using magnetic beads with capture-probe oligonucleotide
target
Sequencing

Figure 4. Principle of long-fragment capture using type IIS restriction enzymes. *Retrieved from Pham et al, Molecular Genetics and Genomics 2016.*

for the PCR-free capture and long-read PacBio sequencing of a 1.1 kbp target spanning the *FMR1* CGG repeat[31]. Importantly, the method is free from amplification-related biases[32–35] and enabled the characterization of native DNA with the possibility to assess its methylation pattern. It is however limited by the high amount of input gDNA required (18 µg), which does not allow to work with low-abundant samples (e.g. clinical biopsies). Also, the success of the approach is restricted to the presence of Type IIS RE sites in close-proximity to the target of interest.

1.2.4 Indirect sequence capture via Xdrop technology

The Xdrop technology (Samplix, Birkerød, Denmark) uses the so-called "indirect sequence capture" (**Figure 5**) to enrich for long fragments (several kbp). High-molecular-weight (HMW) DNA molecules (50-100 kbp) are initially encapsulated in individual droplets, and droplet PCR (dPCR) is used to amplify a detection sequence (DS) of 100–150 bp located near the target of interest. Positive droplets are revealed by staining with a DNA-intercalating dye and are recovered by flow sorting, providing the actual target enrichment. A few hundred target DNA molecules are recovered for multiple displacement amplification after their encapsulation in individual droplets (dMDA) to minimize amplification biases[36–38].

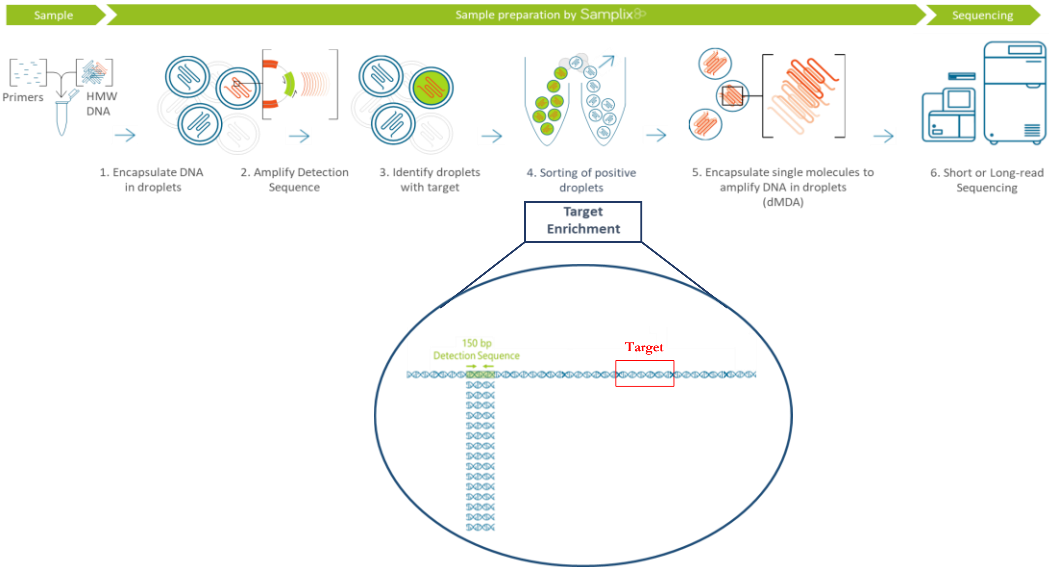

Sample
Sample preparation by Samplix
Sequencing
Primers
HMW DNA
1. Encapsulate DNA in droplets
2. Amplify Detection Sequence
3. Identify droplets with target
4. Sorting of positive droplets
5. Encapsulate single molecules to amplify DNA in droplets (dMDA)
6. Short or Long-read Sequencing
Target Enrichment
150 bp Detection Sequence
Target

Figure 5. Schematic representation of the Xdrop indirect capture workflow. (1) High-molecular-weight DNA molecules are initially encapsulated in individual droplets, and **(2)** droplet PCR (dPCR) is used to amplify a detection sequence of 100–150 bp located near the target of interest. **(3)** Positive droplets are revealed by staining with a DNA-intercalating dye and **(4)** are recovered by flow sorting, providing the actual target enrichment. **(5)** A few hundred target DNA molecules are recovered for

multiple displacement amplification after their encapsulation in individual droplets (dMDA) to minimize amplification biases. **(6)** Amplified target DNA can be sequenced using either short- or long-read sequencing platforms.

The method offers the advantage that the input DNA required to perform the assay is as low as 10-15 ng[36–38]. This is up to 500–1000 times less than the input required for the other long-read capture approaches, mostly based on Cas9-mediated capture or type IIS REs [5–10,31,39–42]. This allows the application of long-read sequencing to limited samples, such as those derived from prenatal testing[49]. To date, the method has been proved useful for the identification of human papilloma virus 18 integration sites in the human genome[36] and CRISPR-Cas9-meidiated off-target genome editing[37]. On a final note, the Xdrop indirect capture has never been applied for the characterization of repeat expansion disorders.

1.2.5 Whole chromosome sorting

Chromosome sorting allows the isolation of entire chromosomes from the genome. Mitotic chromosomes are prepared by blocking cells in mitosis using colchicine. Cells are lysed and filtered out to release the chromosomes. Staining is performed by using DNA-specific fluorochromes to allow chromosomes to be classified according to fluorescence intensity (relative DNA content). Flow sorting is finally employed to generate a flow karyotype, i.e. a distributions of chromosomal DNA content. Ideally, each chromosome forms a distinct peak on the flow karyotype, whose location is proportional to fluorescence intensity and whose volume is proportional to the frequency of occurrence of that chromosome type.

Chromosome sorting has been used for the specific capture of native human chromosomes Y and 1. The method allowed to recover a significant amount of target DNA (4 – 5 μg) which in turn enabled long-read sequencing on the ONT platform, generating up to ~30x coverage[50,51]. In both papers, two highly continuous chromosomes could assembled and used as reference to call SVs. However, the method is laborious, time consuming and does not allow to enrich for specific portions of the chromosome. Importantly, the method's success is strictly dependent on the ability to obtain intact mitotic chromosomes, which is possible only through cell culturing. Also, similarities in size and relative DNA content between chromosomes may lead to peak overlapping during flow sorting and the impossibility to resolve chromosomes.

1.2.6 Cas9-mediated capture

The CRISPR-Cas9 system. The CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) genes and CRISPR-associated proteins (Cas) system is a prokaryotic immune system that confers resistance to foreign genetic elements providing a form of acquired immunity[52,53]. In this system, exogenous DNA from viruses or plasmids is cut into small fragments and incorporated into a CRISPR locus in a series of short ~20 bp repeats. The loci are transcribed, and transcripts are processed to generate small CRISPR RNAs (crRNA), which guide effector Cas endonucleases to recognize and cut invading DNA, based on sequence complementarity[52,54]. The most well characterized and widely used system is the CRISPR-Cas9, originated from type II CRISPR-Cas systems[55,56] and first described in *Streptococcus pyogenes*[57].

Guide RNAs. The *S. pyogenes* Cas9 uses a guide sequence within an RNA duplex, tracrRNA:crRNA (gRNA, **Figure 6B**), to form base pairs with DNA target sequences, enabling Cas9 to introduce a site-specific double-strand break in the DNA[57,58]. The tracrRNA provides structural support to the crRNA whereas the latter provides target recognition based on complementarity (**Figure 6B**) to the genome. A functional crRNA is composed by a protospacer sequence (20 bp) and a protospacer-adjacent motif (PAM) immediately 3'-downstream the protospacer (**Figure 6A**). The presence of the PAM sequence is essential for target recognition and activation of the Cas9[59,60]. In *S. pyogenes*, the motif consists of a random nucleotide followed by two Guanine (5'-NGG-3') bases. The interaction between the gRNA and the Cas9 forms an active ribonucleoprotein (RNP)[58] which then activates the protein. The complex interrogates the DNA in search of the target sequence associated to the PAM[61]. Upon target recognition, Cas9 cleaves the DNA (double-strand breaks) thanks to the complementary activity of the HNH and RuvC-like nuclease domains[59,60] (**Figure 6B**). The design of a crRNA is relatively simple and requires some basic criteria to be followed, such as GC content (40-80%), the absence of secondary structures or mismatches at the seed sequence (first 11 bases upstream PAM, **Figure 6A**). Also, the design is strictly dependent on the presence of the PAM sequence at the target site. Considering the frequency of "GG" dinucleotides in the human genome (5.21%[62]), there is an expected 161,284,793 NGG PAM sites (1 every 42 bases). This makes the use of the CRISPR-Cas9 system extremely programmable and versatile. Owing to its high programmability and specificity, the CRISPR-Cas9 system has been used in a wide variety of biotechnological applications that involve genome editing, gene therapy and, most recently, targeted sequencing [63–65].

Cas9-Assisted Targeting of CHromosome segments (CATCH). The method employs the specificity of Cas9-RNPs for the targeted excision of a given genomic region (**Figure 6C**). Briefly, cells are embedded in agarose plugs and HMW gDNA is released via protein digestion (Proteinase K-based digestion). In-gel target excision is performed by adding two RNP complexes, one on each side of the target. Each RNPs consists of a gRNA (crRNA + tracrRNA targeting the region of interest) and the Cas9 nuclease. The target is then separated from the rest of the genomic DNA (gDNA) by pulsed field gel electrophoresis (PFGE). DNA is isolated from the gel and sequenced using long- or short-read sequencing. CATCH has been first described by Jiang et al.[66] for target cloning of large in-tact genomic fragments (up to 100 kbp). Recently, Gabrieli and coworkers harnessed CATCH for the targeted capture of a 200 kbp region containing the *BRCA1* gene, followed by WGA and ONT/Illumina sequencing[67]. In general, the method shows flexibility as the WGA step allows to yield sufficient target DNA both for ONT and Illumina sequencing libraries. As a drawback, large amount of input gDNA (2 μg) is required. Currently, Sage Scientific is distributing an automatized system to perform the CATCH protocol, however the system is not largely employed due to the need of freshly isolated cells or nuclei.

Cas9-mediated targeted sequencing. The method employs the specificity of RNPs for the targeted excision of genomic regions and sequencing using the ONT long-read platform (**Figure 6D**). Briefly, starting gDNA is dephosphorylated to prevent downstream unwanted adapter ligation. Two RNP complexes are formed, one on each side of the target, consisting of a gRNA (crRNA + tracrRNA targeting the region of interest) and the Cas9 nuclease. Upon target recognition, RNPs-mediated cleavage originates 5'-phosphorylated blunt-ended DNA. The incorporation of a nontemplated dAMP on the 3´-end of cleaved gDNA (dA-tailing) facilitates the specific ligation of the ONT sequencing adapters containing the motor protein. The final library includes also non-target fragments which, in principle, cannot be sequenced since they lack the 5'-phosphate and cannot be ligated by the ONT adapters. The whole library is loaded into the sequencing platform and the enrichment is provided by the sequencing itself, owing to the fact that only target fragments bear sequencing adapters.

The possibility to provide a PCR-free assay enables the unbiased, comprehensive characterization of complex genomic regions, including epigenetic modifications such as methylation. Potentially there are no limitations on target length, as soon as good quality, HMW gDNA is provided and the gRNA design is performed properly. As shows in recent reports[64,68], Cas9-mediated capture coupled to ONT long-read sequencing has enabled the enrichment of genomic targets up to 20 – 100 kbp in length[64,68]. The principal drawback of the method is the large amount of starting material required, typically 1–10 µg DNA[41,43,46,48,64,67–69], which makes it difficult to work with low-abundant samples, as for example those from pre-natal/pre-implant testing or clinical biopsies.

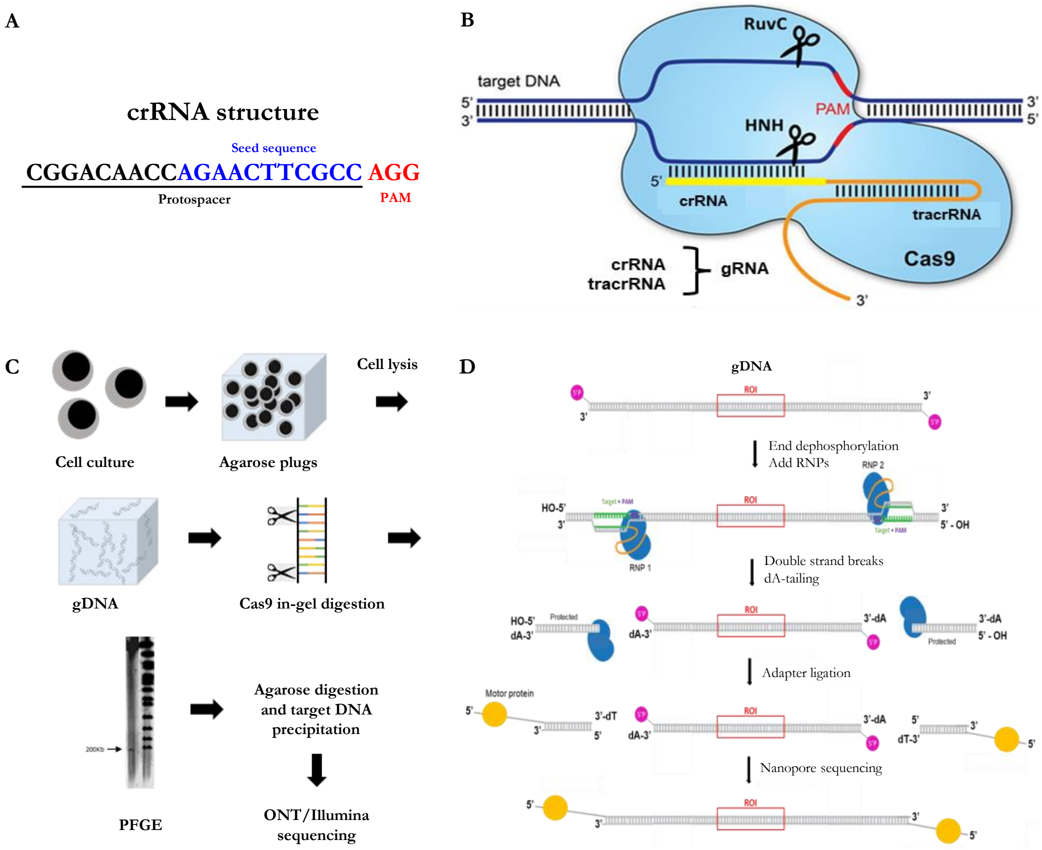


Figure 6. The CRISPR-Cas9 system and its application for targeted sequencing. (A) Structure of a functional crRNA. The protospacer sequence provides target recognition based on complementarity to the genome. For correct target cleavage, the seed sequence (blue) must ne 100% complementary to the target. On the other hand, mismatches will be tolerated toward the 5' end (bold black). The protospacer-adjacent motif (PAM) must always be present 3'-downstream the protospacer and consists of a random nucleotide followed by two Guanine (5'-NGG-3') bases. The PAM sequence is essential to initiate target cleavage upon recognition. Sequences fully complementary to the crRNA but lacking a nearby PAM are ignored by Cas9. **(B)** Structure of a functional ribonucleoprotein complex. The Cas9 nuclease (light blue) forms a complex with the guide RNA. The latter is composed by the tracrRNA (orange) with structural function and the crRNA (yellow) The PAM sequence (red) is found 3'-downstream the protospacer. The Cas9 nuclease domains (HNH and RuvC) are also shown in correspondence of their cleavage sites (~ 3 bp upstream PAM). **(C)** Schematic representation of the CATCH workflow. *Adapted from Gabrieli et al., Nucleic Acids Research 2018.* **(D)** Schematic representation of the Cas9-mediated target sequencing workflow.

1.3 REPEAT EXPANSION DISORDERS

Genomic "dark" regions include tandem repeats, namely sequences of two or more DNA base pairs that are repeated adjacent to each other. The current assembly of the human genome (Hg38) has been reported to contain over one million of annotated tandem repeats[70,71]. They are usually found in non-coding regions of the genome, even though short repetitions of triplets are frequently localized also in coding portions[70]. Their intrinsic repetitive nature makes tandem repeats particularly prone to mutations. They bear indeed the highest mutational rate in the genome and are typically polymorphic and multiallelic, with the longest alleles being the most unstable. However, interruptions of the canonical repeat by alternative repeated motives have been shown to stabilize the expansions[72–75], conferring milder and/or different phenotypes compared to uninterrupted ones[76,77]. Tandem repeat expansions are the causative phenomenon of at least 50 known human disorders[78]. Expansions affecting coding regions usually cause the formation of polyAlanine (PolyA), polyGlutamine (PolyQ) or polyGlycine (polyG) stretches. These may lead either to protein loss of function(polyA) or to protein (polyQ) / peptide (polyG) toxicity[78]. Expansions affecting non-coding portions are usually found either in 5'-untranslated regions (5'-UTRs), introns or 3'-UTRs[79]. Noncoding repeats in 5' regions usually lead to hypermethylation and subsequent gene silencing. Intronic expansions and expansions affecting 3' UTRs are more often linked to RNA toxicity or polypeptide synthesis via repeat-associated non-AUG translation and subsequent formation of cellular/nuclear aggregates. A common characteristic of repeat expansion disorders is genetic anticipation, namely a phenomenon in which the signs and symptoms of some genetic conditions tend to become more severe and/or appear at an earlier age as the disorder is passed from one generation to the next. The mechanism underlining repeat expansion is attributable to the difficulties encountered by the DNA replication machinery within long stretches of repeated DNA. Repeated units seems to be deleted or added to long repetitive tracts as the cellular machinery tries to replicate through DNA hairpins formed by repeated sequences[80].

Genes known to bear repeat expansions include *FMR1* (fragile X mental retardation protein translational regulator 1), *DMPK* (DM1 protein kinase) and *CNBP* (CCHC-type zinc finger nucleic acid binding protein). Expansions in these genes lead to Fragile X syndrome (*FMR1*, FXS), Myotonic dystrophy type 1 (*DMPK*, DM1) and Myotonic dystrophy type 2 (*CNBP*, DM2), respectively (**Table 2**). In particular, FXS repeat expansions have been reported to expand over 600 bp up to 3,000 bp. On the other hand, DM1 and DM2 patients have been reported to bear some of the longest expansions, up to 19.5 kbp[81] (DM1) and 44 kbp[78] (DM2).

Table 2. Schematic representation and features of 3 repeat expansion disorders. *FMR1*, *DMPK* and *CNBP* expansions have been selected as case studies for medium-long, long and ultra-long targets, respectively. *FMR1: fragile X mental retardation protein translational regulator 1, DMPK: DM1 protein kinase, CNBP: CCHC-type zinc finger nucleic acid binding protein.*

		Disease states (# repeats)		
	Expanded motif	Healthy	Pre-mutation	**Mutation**
FMR1 – medium-long expansion	CGG	5 - 53	55 - 200	> 200 up to 1,000
DMPK - Long expansion	CTG	5 - 36	37 - 49	> 50 up to 6,500
CNBP - Ultra-long expansion	CCTG	< 26	27 - 75	75 – 11,000

1.3.1 *FMR1* gene

Repeat expansion in the fragile X mental retardation 1 gene (*FMR1*; MIM# 309550)[82–84] is causative of the Fragile X syndrome (FXS; MIM# 300624), an X-linked disorder characterized by intellectual disability, autism, hyperactivity, long face, large or prominent ears and macroorchidism at puberty and thereafter[85]. The disease is caused by the expansion of CGG trinucleotide repeats in the 5' untranslated region of *FMR1*. The protein encoded by *FMR1* regulates the translation of potentially hundreds of mRNAs, many of which are involved in the development and maintenance of neuronal synaptic connections[86]. The expansion causes the hypermethylation of the promoter and transcriptional silencing[87]. Loss or a shortage (deficiency) of this protein disrupts nervous system functions and leads to the signs and symptoms of FXS. Normal alleles carry 5 – 44 CGG repeats, whereas expanded alleles are classified as intermediate (45 – 54 repeats), pre-mutation (55 – 200 repeats) or full mutation (> 200 repeats). The pre-mutation allele often expands to a full mutation during female germline transmission, thus giving rise to FXS in the progeny. The risk of pre-mutation expansion depends mainly on the number of CGG repeats (with shorter alleles being less likely to expand to a full mutation than larger ones) and the presence of AGG interruptions in the tandem array. Such AGG interruptions increase repeat stability, reduce the risk of expansions[82,88,89] and can modulate the disease phenotype[77,90,91]. Moreover, recent evidence has suggested pronounced repeat variability between individuals and within them (mosaicism) that also modulates the disease phenotype[92,93]. Although much less frequent than microsatellite expansions, intragenic single-nucleotide variants (SNVs) and short insertions or deletions (indels) are significant mutational mechanisms leading to FXS and other repeat-associated diseases[94]. Accordingly, accurate risk prediction in genetic counseling not only requires the precise characterization of repeats, but also the mapping and counting of interruptions within the repeat array and the ability to map additional intragenic variants[95].

1.3.2 DMPK gene

Repeat expansion in the *DMPK* gene (MIM*605377) is causative of Myotonic dystrophy type 1 (DM1; MIM#160900), an autosomal dominant disorder characterized mainly by myotonia, muscular dystrophy, cataracts, hypogonadism, frontal balding, and ECG changes[96,97]. The disease is caused by a (CTG)n repeat expansion in the 3'-UTR of the *DMPK* gene on chromosome 19q13.32[98–100]. The protein encoded by *DMPK* is a serine-threonine kinase which, among others, has the role to inhibit the muscle protein myosin phosphatase, which in turn plays a role in muscle tensing and relaxation[101]. Expanded alleles produce altered mRNAs containing double-stranded hairpin structures, which are not translated and form clumps with other proteins in the cytoplasm. Among these we find Muscleblind Like Splicing

Regulator 1 (*MBNL1*), which is a protein involved in RNA splicing. The depletion of its activity leads to impaired splicing disrupts muscle development and function[102,103]. Non-pathogenic alleles contain up to 36 CTG-repeat units, whereas pre-mutated allele contain between 37 and 49 repeats. In DM1 patients, repetitions range from 50 to up to 6,500 units[81] in congenital forms. Expanded repeats are highly unstable and lead to show genetic anticipation[104]. Moreover, the CTG tract has been shown to expand during an individual's lifetime, resulting in somatic mosaicism and the worsening of the disease symptoms with age[105]. In 3 – 8 % of the DM1 patients the CTG tract of expanded alleles is interrupted by non-CTG tracts such as CCG, CTC or GGC motives [72,106–110]. Importantly, these so-called variant repeats have been reported to stabilize the repeated tract with significant implications in disease onset and progression. As a matter of fact, patients with variant repeats may exhibit delayed onset, unusually mild symptoms, or atypical patterns of symptoms[72,106–110].

1.3.3 CNBP gene

Repeat expansion in the *CNBP* gene (previously *ZNF9,* MIM*116955) is causative of Myotonic dystrophy type 2 (DM2; MIM#602668), an autosomal dominant multisystemic disorder characterized by progressive proximal muscle weakness, myotonia, myalgia, calf hypertrophy and multiorgan involvement with cataract, cardiac conduction defects and endocrine disorders[113,114]. The disease is caused by a (CCTG)n repeat expansion in intron 1 of the *CNBP* gene on chromosome 3q21.3[115]. *CNBP* encodes for a single-stranded DNA-binding protein which is essential for embryonic development in mammals[116]. Nevertheless, still little is known about its function and molecular pathways, though it seems to be involved in cytoplasmic post-transcriptional gene regulatory processes rather than acting as transcription factor[116]. The pathogenic mechanism of DM2 is similar to DM1 and involves the formation of alternative mRNAs containing double-strand hairpin structures which in turn sequester cytoplasmic proteins involved in RNA splicing such as *MBNLs*. This leads to impaired splicing disrupts muscle development and function[103].

The CCTG repeat tract is part of a complex (TG)v(TCTG)w(CCTG)x motif which is generally interrupted in healthy range alleles by one or more GCTG, TCTG or ACTG motifs resulting in repeat stability[117,118,119]. Non-pathogenic alleles contain up to 26 CCTG-repeat units, whereas pre-mutations are made of "pure"$(\mathrm{CCTG})_{<75}$ with still uncertain clinical significance[118,75]. In DM2 patients, repetitions range between ~75 and >11,000 units and represent some of the largest reported so far in repeat expansion disorders[78]. The DM2 mutation shows marked somatic instability and tends to increase in length over time within the same individual, while it does not show a strong bias towards intergenerational expansion and genetic anticipation is rarely seen in DM2 families[115,118,120,121,96]. The few genotype-phenotype studies reported so far in DM2 patients did not reveal any significant associations between the severity of the disease, including the age at onset, and the number of CCTG repeated units[121,122]. Identification of such correlations is indeed strongly challenged by heterogeneity across tissues, somatic instability, and the relative technical difficulty of accurately measuring repeat length in such large microsatellite expansions. For the same reasons, the discovery of additional *cis* genetic modifiers that may protect or exacerbate disease symptoms is hampered.

The genetic features of the *CNBP* microsatellite locus along with its extreme length and high CG-content hampered the ability to sequence expanded alleles in DM2 patients. Indeed, investigators of the original gene-discovery study were unable to sequence the entire CCTG array because of its extremely large size and the high level of somatic mosaicism[115,123,122].

1.3.4 Limitations of the current diagnostic approaches for the analysis of pathological repeat expansions

Most of the known repeat expansion disorders are of recent discovery, probably due to the limitations of the current approaches used in diagnosis (**Table 3**). Best practice guidelines for the diagnosis involve a first step where short-range PCR (SR-PCR) is used to assess whether an individual has two alleles with a low number of repeats. If only one allele is detected, pathogenic expansions can be addressed via long-range PCR (LR-PCR) or repeat-primed PCR (RP-PCR) on genomic DNA (gDNA) or LR-PCR products[120,124–129]. The assay also allows to detect the presence of mosaicism. However, the presence of interruptions may result in aberrant patterns or failure to detect expansions with RP-PCR[106,130]. In such cases, Southern blotting of LR-PCR products or gDNA is usually performed for confirmation[122,125,126]. However, this procedure is time-consuming, requires large amount of DNA and it is not included in the routine workflow of most diagnostic centers. Nonetheless, the combination of these approaches provides clinical sensitivity and specificity approaching 100%[120,131,132].

Despite the high success rate and relatively low costs, these methods are often simultaneously required for reliable diagnosis. Moreover, they can be imprecise when dealing with extremely long expansions, mosaicism and minor alleles[88,133–139]. Above all, they lack single-nucleotide resolution which is crucial for the accurate characterization of both the repeat structure (e.g. interruptions) and its surroundings (e.g. pathogenic SNVs), which have been shown to be involved in disease onset and progression.

NGS approaches provide an unprecedented opportunity for the characterization of repeat expansion at single-nucleotide resolution. However, state-of-the-art approaches rely on short reads which are very accurate but are often limited to normal alleles as the length of the expansions usually exceeds the maximal read length (< 600 bp). During bioinformatic analysis, short reads originating from long tandem repeats typically map to multiple genomic regions, are clipped off, or are discarded. As result, short reads do not allow to accurately determine the repeat length.

Table 3. Overview of current approaches for repeat characterization. The main features of tandem repeats are indicated on the far left. For each feature, the X or ✓ symbols indicate if the approach allows to characterize it or not.

	PCR	Repeat-primed PCR	Southern blot	Short reads
Normal allele	✓	✓	X	✓
Expanded allele	X✓	✓	✓	X
Repeat interruptions	X	✓	X	✓
Mosaicism	✓	✓	✓	X
Methylation	X	X	X	X
Single nucleotide resolution	X	X	X	✓

In this context, the combination of long-read sequencing with long-DNA capture approaches provides a valuable tool for the comprehensive characterization of repeat expansions, which are often associated to pathologic conditions[78]. Long reads can span across the entire repeated region, including their flanking sequences of higher complexity, and provide a single nucleotide resolution. This enables more accurate mapping providing insights on repeat length and structure, and presence of specific features such as interruptions and mosaicism. Moreover, the possibility to map methylation simultaneously to sequencing data acquisition provides an added value to the analysis and the possibility to correlate this molecular feature to the pathological phenotype. Overall, a more accurate characterization of repeat-features would lead to a more precise genotype-phenotype correlations for repeat expansion disorders, that is still lacking/limited for those disease with large microsatellite expansions, such as DM1 and DM2, probably due to the limitations of current diagnostic approaches.

In recent reports, the consistent benefits of long-DNA capture approaches coupled to long-read sequencing technologies have been demonstrated for the characterization of short tandem repeats in different disorders[31,39,48,95,140–143,40–47]. These approaches demonstrated the possibility to sequence DNA fragments several kbp in length, facilitating the accurate genotyping of repeat expansion alleles, with future potential application in diagnostics. In some of these reports, *FMR1*, *DMPK* and *HTT* repeats were amplified by PCR for PacBio long-read sequencing[95,142,143]. PCR may be unsuitable for comprehensive repeat characterization due to amplification-related biases and the difficulties to amplify regions with high CG content[32–35,144]. Also, the use of PCR-based approaches in patients heterozygous for normal and large expansion alleles may lead to the amplification of only the normal allele[145], and polymorphisms surrounding the repeat region may lead to allele bias, dropout, or the misinterpretation of results[135]. Targeted and PCR-free approaches coupled to long-read sequencing have been already reported for the in-depth characterization of repeat expansions in different disorders such as FXS, Frontotemporal lobar degeneration/motor neuron disease, Fuchs endothelial corneal dystrophy, Huntington's disease, Benign adult familial myoclonic epilepsy and Neuronal intranuclear inclusion disease[39–48].

However, the possibility to develop a robust and all-in-one assay for the full characterization of repeat-linked disorders in the clinical setting requires a careful evaluation of these approaches, both in terms of performances and cost-effectiveness. In contrast, a deep benchmarking of targeted-long read sequencing methods, enabling the analysis of clinically relevant repeats, is still lacking.

2. AIM OF THE THESIS

Several approaches for long-DNA-fragment enrichment are available, however their application is limited to the research setting and an extensive benchmarking of their performances is still absent. In this thesis, we assessed and compared the performances of long-DNA capture approaches coupled to long-read sequencing for the characterization of *FMR1*, *DMPK*, and *CNBP* repeat expansions, selected as case studies for medium (*FMR1*), long (*DMPK*) and extremely long (*CNBP*) expanded microsatellites. At this aim, we benchmarked a total of three approaches: indirect sequence capture (Samplix's Xdrop), Cas9-mediated targeted sequencing, and a set of three hybridization-capture methods (PNA, dCas9 and dsDNA-probes). Xdrop's indirect capture was selected owing to its versatility and cheap assay design, although its implementation is still very limited. Cas9-mediated capture is the most commonly utilized approach in combination with long-read sequencing to characterize not only repeat expansions but also other genomic features, such as SNVs, large SVs and CpG methylation. Finally, hybridization-based capture approaches potentially allow to significantly reduce costs but their combination with long-read sequencing is still largely unexplored. The performances of each method were assessed by comparing **I)** type and amount of input gDNA, **II)** enrichment efficiency, **III)** length of enriched DNA fragments and **IV)** additional features, such as costs and need of peculiar instruments.

3. MATERIALS & METHODS

3.1 DNA SAMPLES

Genomic DNA (NA12878, NA06891, NA07537 and NA20241, representing cells with diverse *FMR1* alleles) was purchased from the Coriell Institute for Medical Research. HMW gDNA from the HEK293 cell line was extracted starting from cell pellets (~1 million) and using the Nanobind CBB Big DNA *HMW* kit (Circulomics, now PacBio, Menlo Park, California, United Stat) based on manufacturer's instructions. All the other samples for Xdrop indirect capture were isolated from the whole blood of unrelated healthy donors Blood Center, Verona Hospital) following informed written consent. Samples were de-identified immediately after collection. The study was approved by the Ethics Committee for Clinical Research of Verona and Rovigo Provinces and all the investigations were conducted according to the Declaration of Helsinki. Genomic DNA was extracted from venous blood (0.2 – 0.5 ml) collected in EDTA tubes using the Genomic Tip 100/G kit (Qiagen, Hilden, Germany), NucleoSpin Blood Mini kit (Macherey-Nagel, Düren, Germany) or the Nanobind CBB Big DNA Kit (Circulomics). The Genomic Tip 100/G and NucleoSpin Blood Mini protocols were carried out according to the manufacturer's instructions. For the Nanobind CBB Big DNA kit, we used either the HMW or ultra-HMW protocols with some modifications for 0.5 ml whole blood input. In particular, the volume of all reagents used was increased 2.5-fold and pulse-vortexing was doubled from 10 to 20 times.

Venous blood samples from DM2 patients (N=9) was kindly provided by the Medical Genetics Section of Policlinico Tor Vergata. Enrollment of participants was approved by the institutional review board of Policlinico Tor Vergata (document no. 232/19). All experimental procedures were carried out according to the Declaration of Helsinki. Informed consent was obtained from all the participants and samples were de-identified immediately after collection. Genomic DNA was extracted from 0.5 ml of venous blood collected in EDTA tubes using the Nanobind CBB Big DNA HMW protocol (Circulomics). DNA quantity was measured using the QuBit fluoromether (ThermoFisher Scientific, Waltham, Massachusetts, USA) in combination with the Qubit dsDNA BR Assay Kit (ThermoFisher Scientific). DNA fragment size was assessed by capillary electrophoresis using TapeStation 4150 in combination with the

Genomic DNA ScreenTape assay (both from Agilent, Santa Clara, California, USA) or via Pulsed-filed gel electrophoresis. DNA purity was determined using a Nanodrop spectrophotometer (ThermoFisher Scientific).

3.2 PULSED-FIELD GEL ELECTROPHORESIS

PFGE was run on the CHEF Mapper electrophoresis system (Bio-Rad Laboratories, Hercules, California, USA). Seven hundred ng DNA was resolved by PFGE using a 1% agarose gel that was let solidified overnight at RT for 3 h.

The electrophoresis chamber was filled by 2.2 l of 0.5x Tris-Borate-EDTA (TBE) buffer. The run was set as follows and according to according to the expected size of extracted DNA:

- From 250 – 2,200 kbp with a Two State Mode (24 h) that consist of two field vectors, with each vector having the same voltage and duration but separated in direction by a 120° definable included angle, with an Initial Switch Time of 60" and a final switch time of 90".
- From 50 – 1,000 kbp with a Two State Mode (20 h) that consist of two field vectors, with each vector having the same voltage and duration but separated in direction by a 120° definable included angle, with an Initial Switch Time of 35" and a final switch time of 90".
- From 5 - 450 kbp with a Two State Mode (20 h) that consist of two field vectors, with each vector having the same voltage and duration but separated in direction by a 120° definable included angle, with an Initial Switch Time of 5" and a final switch time of 35".

Three λ DNA markers (either liquid or embedded in agarose) were chosen for PFGE run differing in their size range: CHEF DNA Size Standard (size range <8.3 – 48.5 <kbp, Bio-Rad), MidRange PFG Marker (size range <15 – 291 <kbp, New England Biolabs), Lambda PFG Ladder (size range <48.5 – 1018 <kbp, New England Biolabs) and CHEF DNA Size Marker (size range <225 – 2200 <kbp, Bio-Rad). After the run, the gel was stained for 30' on a Hula mixer at 80 rpm in 400 ml of 0.5x TBE buffer supplemented with 40 µl Syber Gold (Thermo Fisher Scientific). Subsequently the gel was washed for 30' minutes with fresh 0.5x TBE buffer. Gel imaging was performed on a ChemiDoc Touch Imaging System with Image Lab Touch Software (Bio-Rad).

3.3 XDROP INDIRECT CAPTURE

3.3.1 Droplet generation and dPCR

Before enrichment, DNA samples were purified using 1x HighPrep MagBio beads (MagBio Genomics, Gaithersburg, MD, USA) and diluted with DNase-free water to 5 ng/µl. Detection sequence-specific primers for *FMR1, DMPK* and *CNBP* enrichment were designed using the Samplix primer design tool (https://samplix.com/primer, Table 4)). The dPCR reaction consisted of 20 µl 2x dPCR mix (Samplix), 0.8 µl of each primer (10 mM), 2 µl 5 ng/ µl DNA, and water to 40 µl. Droplets were generated using a dPCR cartridge and Xdrop droplet generator (both from Samplix). Droplets were then transferred to four tubes and dPCR was carried out by heating to 94°C for 2' followed by 40 cycles of 94°C for 3" and 60°C for 30" at a ramping rate of 1.5 °C/s.

Table 4. Primer pairs used in this thesis. For each primer is shown the reference assay, the target gene, the genomic coordinates, the expected amplicon length and the efficiency as determined via qPCR and using a calibration curve.

Assay	ID	Primer	Sequence	Chromosome	[Start]	[End]	Amplicon length (bp)	Efficiency
Xdrop indirect capture	FMR1_dPCR	Fw	GAGCCCTAGTCCTCACCCAAT	X	147,917,927	147,917,947	140	NA
		Rev	CCCTACCTATCAGGCAAAGCT		147,918,046	147,918,066		
	FMR1_qPCR	Fw	TCATTGGTGGTCGGGTGTAC		147,917,152	147,917,171	111	1.08
		Rev	AGCGACACCTCACATTCCTT		147,917,243	147,917,262		
	DMPK_dPCR	Fw	AGTCTAGGTCACTGCTGGGT	19	45,777,164	45,777,183	137	1.03
		Rev	GGACCTTCCTTGCTGAGTCA		45,777,049	45,777,068		
	DMPK_qPCR	Fw	TCTCCCCGTCCAGATATAGG		45,771,261	45,771,280	216	0.82
		Rev	AGAGAGAAGGGACAGGTGAC		45,771,065	45,771,084		
	CNBP_dPCR	Fw	GCTTGGAGAGGTGAGCACAT	3	129,171,389	129,171,408	157	NA
		Rev	GCCACCTCTACCGCAGTTAT		129,171,252	129,171,271		
	CNBP_qPCR	Fw	GGCGTTTTGTTCTGCATGGT		129,170,637	129,170,656	124	1.06
		Rev	TGCTGCAGTTGATGGCTACA		129,170,533	129,170,552		
Cas9-mediated capture	DMPK_gRNA1	Fw	CTGAAGTGGCAGTTCCAGCG	19	45,771,908	45,771,927	275	1.03
		Rev	CTCGCGTAGTTGACTGTGGG		45,771,653	45,771,672		
	DMPK_gRNA2	Fw	CACCGAGCGTCGAGAAGAGG		45,769,462	45,769,481	132	0.88
		Rev	GGCTAGAAAGTTTGCAGCAACTT		45,769,350	45,769,372		
	DMPK_gRNA3	Fw	AGGGTGTGTCAGGTGGATGAG		45,779,108	45,779,128	187	1.02
		Rev	CCCAGACACTCCTTCCCTAG		45,778,942	45,778,961		
	DMPK_gRNA4	Fw	AGACGCAGGGACAGGAATGG		45,765,441	45,765,460	276	0.97
		Rev	GTTCCGCTTACAGCTAGTACC		45,765,185	45,765,205		
	DMPK_gRNA5	Fw	TAAGTAAGGGTGTGTGTGTTGC		45,772,440	45,772,461	155	0.95
		Rev	AACAAATCAGGATTCCCACCTG		45,772,307	45,772,328		
	CNBP_gRNA1	Fw	GACAAGTCTATCACAAGGGACC	3	129,175,970	129,175,991	132	1.19
		Rev	TAAAAGAGGTCTTTGAGTGGCC		129,175,860	129,175,881		
	CNBP_gRNA2	Fw	CAAAGATCTGACTGCAGCCATG		129,171,755	129,171,776	153	0.82
		Rev	GACAAAATACCTCTATCCGAGG		129,171,624	129,171,645		
/	PRL	Fw	AGGGAAACGAATGCCTGATT	6	22,297,096	22,297,115	120	1.02
		Rev	GCAGGAAACACACTTCACCA		22,296,996	22,297,015		

3.3.2 Positive droplet sorting

Following dPCR, droplets were collected in a single tube, diluted with 1 ml dPCR buffer (Samplix) and stained with 10 ml droplet dye (Samplix). Droplets were sorted on a FACS Aria Fusion II (Becton Dickinson, Franklin Lakes, NJ, USA), with instrument settings adjusted to FSC = 210, SSC = 250 and FL1 = 370. The positive droplets were gated on FL1 fluorescence and the sorting mode was set to "Yield". Sorted droplets were collected in 15 ml water.

3.3.3 dMDA

Sorted droplets were mixed with 20 µl Break solution and 2 µl Break color (Samplix), and 10 µl of the resulting aqueous phase was used as a template for dMDA. The reaction mix consisted of 4 µl dMDA buffer, 1 µl dMDA enzyme, 10 µl template, and water to 20 µl. Droplets were generated as above, while running the dMDA program. Afterwards, the droplets were incubated for 16 h at 30°C (lid at 75°C) followed by 10' at 65°C to terminate the reaction. The dMDA droplets were broken using 20 µl Break solution and 1 µl Break color as above.

3.3.4 qPCR analysis

Total DNA released from dMDA droplets was quantified using a Qubit fluorimeter and the Qubit HS DNA quantification kit (Thermo Fisher Scientific). The size range of the amplified DNA was analyzed on a TapeStation 4150 using the Genomic DNA ScreenTape assay (both from Agilent Technologies). Fold enrichment of target DNA was assessed by qPCR using the KAPA library Quant qPCR mix (Roche, Basel, Switzerland), 10 ng DNA, and 10 µM each of forward and reverse validation primers (**Table 4**). The qPCR reaction was performed on a QuantStudio™ 3 Real-Time PCR System (ThermoFisher Scientific) with the following program: initial denaturation at 94°C for 2' followed by 40 cycles of 94°C for 3" and 60°C for 30". Fold enrichment was determined using an online calculator (https://samplix.com/calculations). Usually, samples with ≥ 100-fold enrichment at qPCR showed also robust enrichment and breath of coverage after sequencing and thus were selected for downstream analysis.

3.3.5 ONT sequencing of enriched DNA

We sequenced 1–1.5 µg of the enriched DNA samples from the Xdrop workflow using the ONT platform, pooling two replicates when necessary. Amplified DNA was initially debranched using 15 units of T7 endonuclease I in 30 µl for 15'. Debranched DNA fragments were isolated by size selection using AMPure XP beads (Beckman Coulter, High Wycombe, UK) in the presence of 15% polyethylene glycol (Sigma–Aldrich, St Louis, MO, USA). The ONT sequencing library was generated using the Ligation sequencing Kit SQK-LSK109 (ONT, Oxford, UK) according to the manufacturer's instructions with minor modifications. Briefly, DNA was end-repaired using the NEBNext FFPE DNA Repair Mix (New England Biolabs, Ipswich, MA, USA) at 20°C for 10’, and subsequently end-prepped with the NEBNext End repair/dA-tailing Module (New England Biolabs) at 20°C for 20’. Sequencing adapters were ligated at room temperature for 10’. Finally, the 30–50 fmol library was loaded onto a MinION R9.4.1 flow-cell (ONT) and standard settings were applied for a run time of ~16 h using MinKNOW (ONT, v20.06.5).

3.3.6 ONT sequencing data and repeat analysis

All the analysis were performed by the bioinformaticians of the University of Verona’s Functional genomics lab. Raw ONT fast5 files were base-called using Guppy v4.2.2 in high-accuracy mode. Reads were quality filtered using NanoFilt v2.7.1[146], with a minimum quality score of 7. Reads were then mapped to the hg38 human reference genome using Minimap2 v2.17-r941[147]. The ONT datasets showed a large fraction of bases (59.3%) mapping as supplementary alignments within the same genomic region, but not recurrent at the same position, suggesting the presence of chimeric reads possibly derived from dMDA as previously reported[148,149]. To exploit the full sequencing dataset, both primary and supplementary alignments completely spanning the *FMR1* repetitive region were considered. Repeat length of normal and expanded alleles was determined from consensus sequences obtained by the de novo assembly of the extracted sequences using the CharONT pipeline (https://github.com/MaestSi/CharONT). Sequences assigned to each allele were processed separately. Following two rounds of polishing, consensus sequences for each allele were searched for repeat motifs using Tandem Repeat Finder v4.09[150].

The presence of somatic mosaicism was investigated by aligning reads to sequences flanking the repeat, searching for repeat motifs, and visualizing alignments in a genome browser using the MosaicViewer_FMR1 pipeline (https://github.com/MaestSi/MosaicViewer_FMR1). Alignments were visualized in the IGV genome browser v2.8.3[151].

3.3.7 Illumina sequencing

Amplified DNA was fragmented using a Covaris sonicator to achieve an average size of 400 bp, and Illumina PCR-free libraries were prepared from ~200–400 ng DNA using the KAPA Hyper prep kit and unique dual-indexed adapters (5 µL of a 15 µM stock) according to the supplier’s protocol (Roche). The library concentration and size distribution were assessed on a Bioanalyzer (Agilent Technologies). Barcoded libraries were pooled at equimolar concentrations and sequenced on a NovaSeq6000 instrument (Illumina, San Diego, CA, USA) to generate 150-bp paired-end reads.

3.3.8 Illumina sequencing data analysis

All the analysis were performed by the bioinformaticians of the University of Verona's Functional genomics lab. Illumina fastq files were quality checked using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and low-quality nucleotides and adaptors were trimmed using fastp[152]. Reads were then aligned to the reference human genome version GRCh38/hg38 using BWA-MEM v0.7.17 (https://arxiv.org/abs/1303.3997). All bam files were cleaned by local realignment around indel sites, followed by duplicate marking and recalibration using Genome Analysis Toolkit v3.8.1.6. BamUtil v1.4.14 was used to clip overlapping regions of the bam file in order to avoid counting multiple reads representing the same fragment. CollectHsMetrics by Picard v2.17.10 was used to calculate fold enrichment to determine enrichment quality.

3.4 CAS9-MEDIATED TARGETED SEQUENCING

3.4.1 crRNA design

Design of crRNAs was conducted using the online tool CHOPCHOP (https://chopchop.cbu.uib.no/) following ONT's recommendation (https://community.nanoporetech.com/info_sheets/targeted-amplification-free-dna-sequencing-using-crispr-cas/v/eci_s1014_v1_reve_11dec2018). The upstream crRNA was designed on the +ve strand while the downstream crRNA was designed on the -ve one, making sure that the excised fragment was at least 3 kbp. Candidates were manually checked for unique mapping via alignment on the human genome (Hg38) using BLAST and excluding region overlapping common SNPs (MAF > 0.01, dbSNP database). Selected crRNAs are shown in **Table 5.** The crRNAs were purchased from Integrated DNA Technologies (IDT, Coralville, Iowa, USA) at 2 nmol scale.

Table 5. Guide-RNAs used in this thesis. The design has been performed using the CHOCHOP online tool. Candidates were manually checked for unique mapping via alignment on the human genome (Hg38) using BLAST and for the presence of SNPs (MAF < 0.001) using the dbSNP database. For each gRNA is shown the target gene, the sequence, the PAM sequence, the genomic coordinates, the strand location and the cut efficiency (i.e. % of Cut DNA) as determined via qPCR.

ID	Target	Target sequence	PAM	Chr	[Start]	[End]	Strand	Cut efficiency (qPCR)
gRNA1	DMPK	CGGACAACCAGAACTTCGCC	AGG	19	45,771,773	45,771,792	+	80.3%
gRNA2	DMPK	TCGAGCTTGCGTCCCAGGAG	CGG	19	45,769,376	45,769,395	-	67.6%
gRNA3	DMPK	GGTCTTCAGGAATTCTAACG	GGG	19	45,778,972	45,778,991	+	65.6%
gRNA4	DMPK	AGTCCCCCACGTATATGGCA	GGG	19	45,765,242	45,765,261	-	84.9%
gRNA5	DMPK	TCGATCTTTGAGCTGAGCGC	TGG	19	45,772,405	45,772,424	+	37.4%
gRNA6	DMPK	GTTGGAAGACTGAGTGCCCG	GGG	19	45,770,437	45,770,456	+	NA
gRNA7	DMPK	ATAAATACCGAGGAATGTCG	GGG	19	45,769,783	45,769,802	-	NA
gRNA8	DMPK	TGCGAACCAACGATAGGTGG	GGG	19	45,770,067	45,770,086	-	NA
gRNA1	CNBP	CCACCTGATTCACTGCGATA	GGG	3	129,175,929	129,175,948	+	74.4%
gRNA2	CNBP	GGCTTCTCATTCCACGACCA	CGG	3	129,171,664	129,171,683	-	84.4%

3.4.2 RNP complex assembly

RNA assembly was performed following the ONT protocol (Version ENR_9084_v109_revD_04Dec2018). A mixture of the crRNAs (10 µM each) and the transactivation crRNA (tracrRNA, 10 µM, IDT) in duplex buffer was denatured at 95°C for 5' and cooled to room temperature for 10' to form crRNA:tracrRNA duplexes. RNPs were formed by mixing 10 µM gRNAs with 62 µM Alt-R® S.p. HiFi Cas9 Nuclease V3 (IDT) in 1X CutSmart Buffer (New England Biolabs) and incubating 30' at RT.

3.4.3 qPCR analysis

Primer pairs flanking each gRNA cut site were manually designed and are reported in **Table 4.** Two µg of HMW gDNA from the HEK293 cell line were directly subjected to Cas9-mediated cleavage with the selected gRNAs as previously described, ending the protocol right after the cut reaction. Then, 5 ng of intact and cleaved DNA were amplified using the primer pair of interest (10 µM each), separately, and KAPA library Quant qPCR mix (Roche). Both cleaved and intact DNA were also amplified using primers for the Prolactin gene (10 mM each) that was used as internal control unaffected by cutting. The qPCR reaction was performed on a QuantStudio™ 3 Real-Time PCR System (ThermoFisher Scientific) with the following program: initial denaturation at 95°C for 10' followed by 40 cycles of 95°C for 15" and 60°C for 1'. Fold change (i.e. amount of uncut DNA) was obtained using the following formula which allowed to normalize the results on the efficiency of the primers:

$$\textit{Fold change} = \frac{Eff.\ Target\ gene\ \hat{}\ (Ct\ Uncut\ DNA - Ct\ Cut\ DNA)}{Eff.\ Control\ gene\ \hat{}\ (Ct\ Control\ Uncut\ DNA - Ct\ Control\ Cut\ DNA)}$$

Cut efficiency (i.e. % of cut DNA) was then determined using the following formula:

$$\textit{Cut efficiency} = (1 - Fold\ change)\ x\ 100$$

3.4.4 Cas9-mediated capture

In order to preserve DNA integrity, all mixing steps were performed via tube flicking and, when strictly necessary, using wide bore pipette tips. Input gDNA, 1 – 10 μg, was dephosphorylated using 15 units of Quick Calf Intestinal Phosphatase (New England Biolabs) in 1x CutSmart Buffer (New England Biolabs) for 10' at 37°C and then 2' at 80°C for enzyme inactivation. Next, 10 ul of pre-assembled RNPs were mixed to dephosphorylated DNA for target cleavage and simultaneous dA-tailing with dATP, using 5 units of Taq DNA polymerase (New England Biolabs) and 10 mM dATP (New England Biolabs). Cas9-meidated digestion and dA-tailing were performed 20' at 37°C and 5' at 72°C to inactivate the Cas9. Five μl of ONT's sequencing adapters (AMX, ONT) were ligated for 10' at RT to the cleaved, dA-tailed DNA using 20 μl Ligation Buffer (ONT) and 10 μl of Quick T4 DNA ligase from the NEBNext® Quick Ligation Module (New England Biolabs). The reaction was stopped by adding 1 volume of 10 mM Tris-EDTA pH 8. Then, the mixture was purified using 0.3x AMPure XP magnetic beads (Beckman-Coulter). Beads were washed twice using 250 μl of Long Fragment Buffer (ONT) and DNA was eluted 10' at RT using 13 μl of Elution Buffer (ONT).

3.4.5 Multiplexed Cas9-mediated capture

Some of the samples were processed in multiplex using the Cas9-mediated PCR-free enrichment native barcoding protocol provided by ONT in combination with the native barcoding kit from ONT (EXP-NBD104). Briefly, one mix per each barcode was prepared by mixing 3 µl of the ONT native barcode with 5 µl of nuclease-free water and 50 µl of Blunt/TA Ligase Master mix (New England Biolabs). The mix was added to each cleaved and dA-tailed DNA and incubated 10' at RT. Barcoded samples were purified using 0.5x AMPure XP magnetic beads (Beckman-Coulter) and eluted in 14 µl of nuclease-free water. All available nanograms from each sample were pooled in a final volume of 65 µl nuclease-free water. Five µl of ONT's sequencing adapters (AMXII from EXP-NBD104, ONT) were ligated for 10' at RT to the cleaved, dA-tailed DNA using 20 ul Ligation Buffer (ONT) and 10 µl of Quick T4 DNA ligase from the NEBNext® Quick Ligation Module (New England Biolabs). The reaction was stopped by adding 1 volume of 10mM Tris-EDTA pH 8. Then, the mixture was purified using 0.3x AMPure XP magnetic beads (Beckman-Coulter). Beads were washed twice using 250 µl of Long Fragment Buffer (ONT) and DNA was eluted 10' at RT using 13 µl of Elution Buffer (ONT).

3.4.6 ONT Sequencing

Purified DNA was mixed with 37.5 µl of Sequencing buffer (ONT) and 25.5 µl of Library loading beads (ONT). Library was loaded on a FLO-MIN106D (R9.4.1) flow cell and sequenced using MinKNOW (ONT, v20.06.5) until plateau was reached.

3.4.7 ONT Sequencing data and repeat analysis

All the analysis were performed by the bioinformaticians of the University of Verona's Functional genomics lab. Raw ONT fast5 files were base-called using Guppy v3.4.5 with high-accuracy mode. Reads from multiplexed runs were demultiplexed with Guppy v3.4.5. Reads were quality filtered with NanoFilt v2.7.1[146], requiring a minimum quality score of 7. Reads spanning the full repeat were identified via *in-silico* PCR and using 100 bp primers placed on the repeat's flanking regions (minimum alignment identity: 80%). These "complete sequences" were extracted and used for the subsequent analyses. Complete sequences were assigned either to the wild-type or to the expanded allele based on their length. Reads from each allele and each sample were then processed separately. An accurate consensus sequence was obtained collapsing reads from the wild-type allele using the CharONT pipeline (https://github.com/MaestSi/CharONT) as described for the Xdrop workflow. The polished consensus sequences were searched for repeats using Tandem Repeat Finder v4.09[150]. Reads from the expanded allele were aligned to sequences flanking the repeat, searched for repeats with motif "TG", "CCTG" and "TCTG" , and visualized using the IGV genome browser v2.8.3[151]. For ease of viewing, reads coordinates corresponding to an annotated repeat were replaced by a single nucleotide stretch of length equalling the annotated repeat and assigned a specific colour. Scripts for annotating repeats and generating simplified reads for the expanded allele are reported in https://github.com/MaestSi/MosaicViewer_CNBP.

3.5 HYBRIDIZATION-BASED APPROACHES

3.5.1 PNA design

PNAs were designed using the online PNA-Bio tool (https://www.pnabio.com/support/PNA_Tool.htm) and were chosen based on length (12 - 21 bp) self-complementarity (score=0), purine content (< 60%), and short purine stretches (especially G, < 6). Sequences were then manually checked for unique mapping via alignment on the human genome (Hg38) using BLAST and excluding region overlapping common SNPs (MAF > 0.01, dbSNP database). Selected PNA was 5' - TCT CCG CCC AGC TCC AGT CC – 3' with genomic coordinates chr19 : 45,770,358 - 45,770,377 and was located 93 bp upstream the *DMPK* microsatellite. Biotinylated PNA was purchased from Panagene Inc. (Yuseong-gu, Daejeon, South Korea) at 25 nmol scale. Once delivered, the PNA was resuspended H_2O to a final concentration of 5 µM.

3.5.2 PNA-mediated capture

Five µg HMW gDNA in 97 µl Tris buffered saline 25 mM Tris (pH 7.2) - 150 mM NaCl Sodium Phosphate Buffer pH 7.5 – EDTA 1 mM were mixed with 3 µl PNA (15 pmol), incubated 10' at 95°C and then cooled at RT for 10' for PNA-DNA hybridization. Ten µl of Dynabeads™ M-270 Streptavidin beads (0.1 mg, ThermoFisher Scientific) were washed three times using 100 ul of wash buffer (25 mM Tris pH 7.2 - 150 mM NaCl) on a magnetic stand. Beads were then resuspended in 10 ul wash buffer. Washed beads were added to the hybridization mix (PNA-DNA triplexes) which was then incubated 30' at RT in a rotator mixer to prevent beads from precipitating. Following hybridization, beads were placed on magnetic stand and the surnatant was removed and discarded. Washing steps (or no washes) were performed according to **Table 12**. Following the final wash, beads were concentrated on magnetic stand and the surnatant was discarded. Beads were gently resuspended in 10 µl H_2O and were heated at 95°C for 5'. Then, beads were quickly concentrated on a magnetic stand and the eluate was transferred on a collection tube.

3.5.3 dCas9-based capture

Dead-Cas9-3XFLAG™-Biotin Protein was purchased from MilliporeSigma (Burlington, Massachusetts, USA). Once delivered, the enzyme was resuspended using the provided dilution buffer to a final concentration of 62 µM. The gRNAs (6, 7 and 8) were designed as previously described and are reported in **Table 5**. The gRNAs and the RNP complexes were prepared as previously described and using 0.3 µl of biotinylated dCas9 (62 µM) instead of the canonical Cas9. Five µg HMW gDNA in 37 µl 1X CutSmart Buffer (New England Biolabs) were combined with 5 µl of dCas9 RNPs and then incubated 10' at 37°C. Ten µl of Dynabeads™ M-270 Streptavidin beads (0.1 mg, ThermoFisher Scientific) were washed three times using 100 ul of wash buffer (25 mM Tris pH 7.2 - 150 mM NaCl) on a magnetic stand. Beads were then resuspended in 10 ul wash buffer. Washed beads were added to the hybridization mix (dCas9-DNA complexes) which was then incubated 30' at RT in a rotator mixer to prevent beads from precipitating. Following hybridization, beads were placed on magnetic stand and the surnatant was removed and discarded. Beads were directly resuspended in 10 µl H_2O and were heated at 95°C for 5' (this step also denatures dCas9). Then, beads were quickly concentrated on a magnetic stand and the eluate was transferred on a collection tube.

3.5.4 DNA-probe based capture

Biotinylated dsDNA-probes (120 bp each) were designed to capture a 1.9 kpb region spanning the *DMPK* microsatellite and were purchased from Twist biosciences (South San Francisco, California, USA). The hybridization capture protocol was adapted from the Appendix 10 of the Twist Human Core Exome Enrichment protocol (Twist bioscience) with some modifications and starting from 5 µg of HMW gDNA in in 12 µl Tris-HCl pH 8. Hybridization was performed for 16 h at 70°C following a denaturation step at 95°C for 5', in the presence of 30 µl Hybridization enhancer. Streptavidin-based pull down and washing steps were performed based on manufacturer’s instructions. Following the final wash, no PCR was performed and beads were concentrated on magnetic stand and the surnatant was discarded. Beads were gently resuspended in 20 µl H_2O and were heated at 95°C for 5’. Then, beads were quickly concentrated on a magnetic stand and the eluate was transferred on a collection tube.

4. RESULTS

4.1 REPEAT CHARACTERIZATION BY XDROP INDIRECT CAPTURE

We first tested the Xdrop's indirect capture for the enrichment of long-DNA fragments spanning the *FMR1*, *DMPK* and *CNBP* microsatellites. We selected this method as first one to test because the assay design is the less expensive and most versatile, depending on a single primer pair. In addition, the method can be potentially coupled with both ONT long-read and Illumina short-read sequencing.

4.1.1 FMR1, DMPK and CNBP microsatellite enrichment using the Xdrop technology

Specific primer pairs were designed to amplify a Detection Site (DS) by droplet PCR (dPCR) at few kbp (< 5kbp) from the microsatellite repeat on the *FMR1*, *DMPK* and *CNBP* genes (**Figure 7A**, **B** and **C**). Additional primer pairs were designed to monitor enrichment by qPCR (Figure **7A**, **B** and **C**).

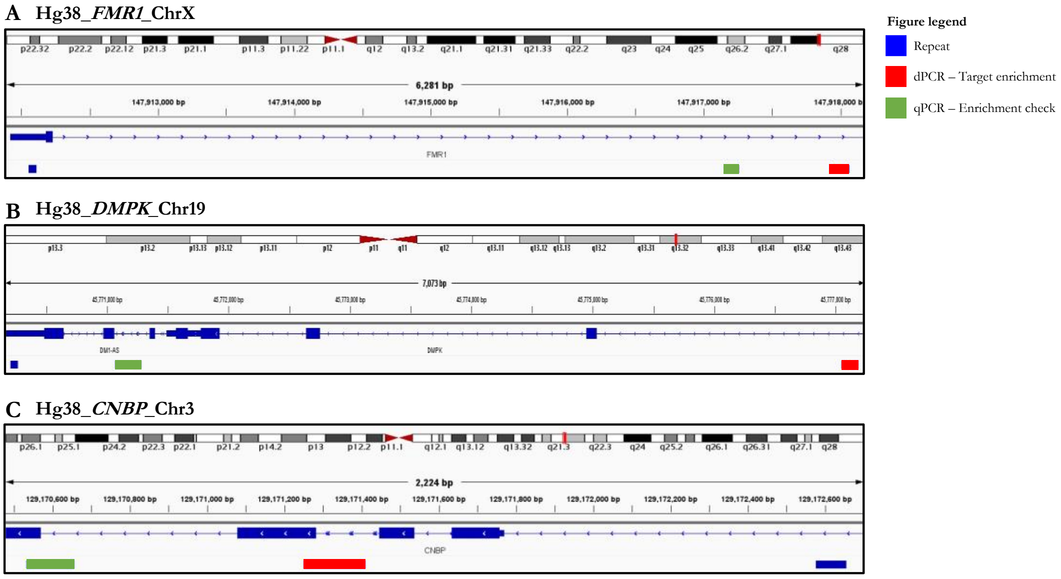


Figure 7. Localization of dPCR and qPCR primer pairs at the *FMR1*, *DMPK* and *CNBP* locus. Integrative Genomics Viewer (IGV) visualization of the **(A)** *FMR1* locus on the X chromosome, **(B)** *DMPK* locus on chromosome 19 and the **(C)** *CNBP* locus on chromosome 3. The figure shows the localization of the microsatellite (blue), primers used to amplify the detection sequence by dPCR (red), and to assess the enrichment by qPCR (green) after applying the Xdrop workflow.

The three Xdrop assays were tested in parallel on samples containing gDNA fragments > 60 kbp, extracted from healthy control donors, Coriell reference samples and a cell line (HEK293). Following Xdrop-mediated encapsulation and dPCR, a clear cloud of positive droplets was visible by FACS analysis for all targets except *DMPK*, where a positive cloud was indeed present but not as defined as *FMR1* and *CNBP* ones (**Figure 8A**, **B** and **C**). We sorted an average of 462, 230 and 424 positive droplets, allowing the recovery of 1.3, 0.9 and 0.8 ug of enriched DNA after dMDA for *FMR1*, *DMPK* and *CNBP* targets, respectively (**Figure 8D-E**). Based on qPCR analysis, average on-target enrichment was 319-fold, 65-fold and 127-fold for *FMR1*, *DMPK* and *CNBP* microsatellites, respectively (**Figure 8F**). *DMPK* and *CNBP* showed a lower enrichment and DNA recovery than *FMR1*, consistently with the fact that they showed a positive cloud that was either not very focused or not well-separated from the background signal. The enriched DNA obtained post-dMDA amplification was 9-10 kbp in length (**Figure 8G, H and I**). In the case of the *FMR1* microsatellite, such DNA length was more than enough to span full-mutation alleles, ranging from >600 bp to 2.2 kbp[95]. In contrast, since *DMPK* and *CNBP* microsatellites are known to expand up to 19.5 kpb[81] and 44 kbp[78], respectively, we concluded on the basis of such results that the Xdrop approach was either highly-risky or unsuitable for the characterization of these repeat expansions. For this reason, further experiments have been performed only on the *FMR1* microsatellite.

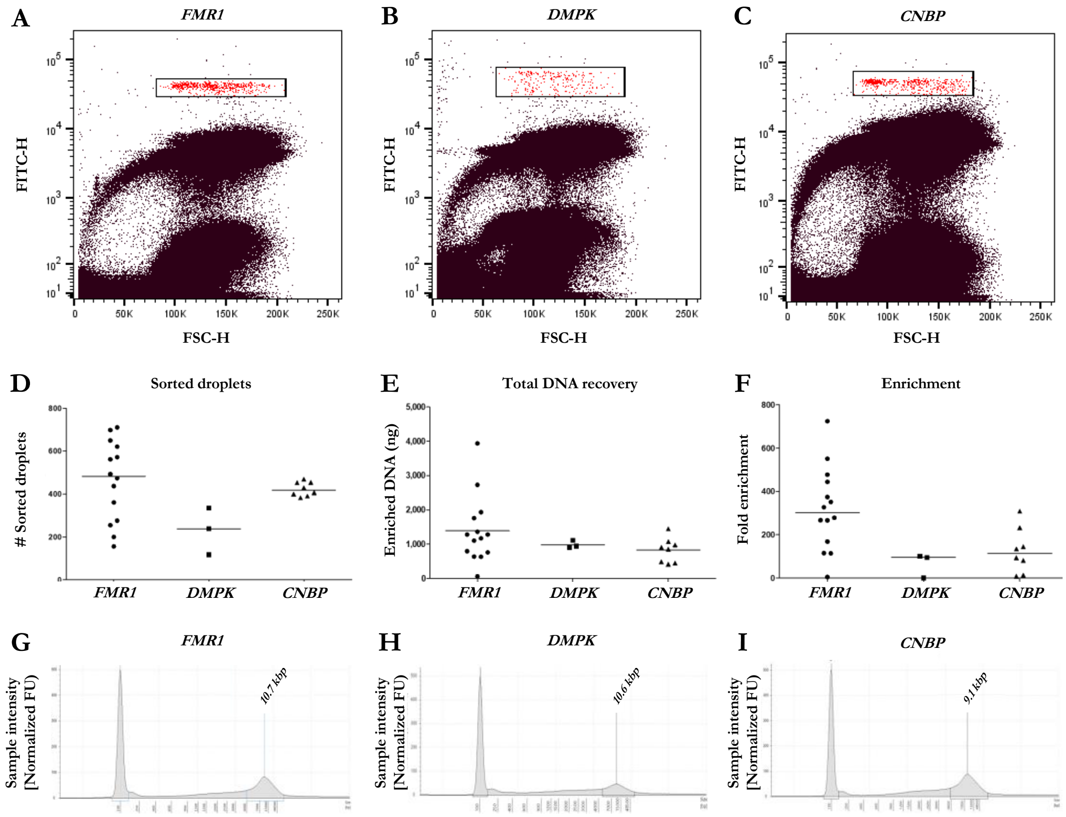


Figure 8. Statistics of *FMR1*, *DMPK* and *CNBP* enrichment using the Xdrop technology. FACS dot plots showing forward scatter (FSC-H) *vs* fluorescence intensity (FITC-H) of droplets obtained after the dPCR step. The gate (red events) identifies the positive droplets that are sorted for **(A)** *FMR1* **(B)** *DMPK* and **(C)** *CNBP* targets, respectively. **(D)** Number of positive sorted droplets, **(E)** quantity of amplified DNA recovered after dMDA, **(F)** fold enrichment of *FMR1*, *DMPK* and *CNBP* determined by qPCR after applying the Xdrop workflow. Fragment distribution of dMDA target DNA samples obtained by capillary electrophoresis of **(G)** *FMR1* **(H)** *DMPK* and **(I)** *CNBP* targets, respectively.

Target enrichment was tested across five gDNA extraction methods (**Figure 9A** and **Figure 10A**) yielding from standard to Ultra-HMW gDNA fragments. These experiments aimed at determining if Xdrop could be influenced either by the DNA integrity (i.e. viscosity) or by the carryover of extraction-method-specific contaminants. Following flow sorting, a clear cloud of positive droplets (531 on average) was visible for all methods (**Figure 10B**), allowing the recovery of 1.1 µg of enriched DNA after dMDA (**Figure 9A, B**). Average on-target enrichment was 351-fold across all methods based on qPCR analysis (**Figure 9C**). Although the Circulomics ultra-HMW protocol resulted in highly variable enrichments, no significant differences were observed among the extraction methods on average, with the exception of Qiagen columns (which did not achieve successful enrichment).

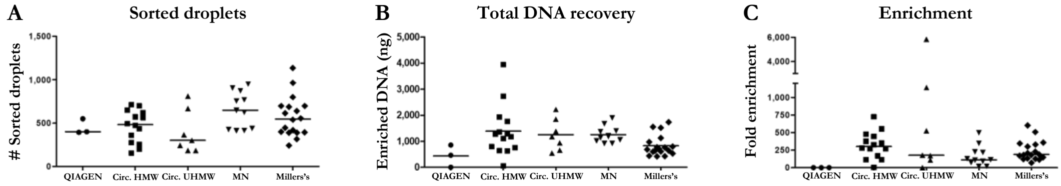


Figure 9. Statistics of FMR1 enrichment using the Xdrop technology. (A) Number of positive sorted droplets, **(B)** quantity of amplified DNA recovered after dMDA, **(C)** fold enrichment of FMR1 determined by qPCR after applying the Xdrop workflow to DNA samples extracted with different methods: Genomic Tip kit (Qiagen, N=3), Circulomics Nanobind CBB Big DNA Kit using either the HMW protocol (Circ. HMW, N=14) or the ultra-HMW protocol (Circ. UHMW, N=7), the NucleoSpin Blood kit (Macherey-Nagel, MN, N=11), or Miller's protocol (Coriell samples, N=18).

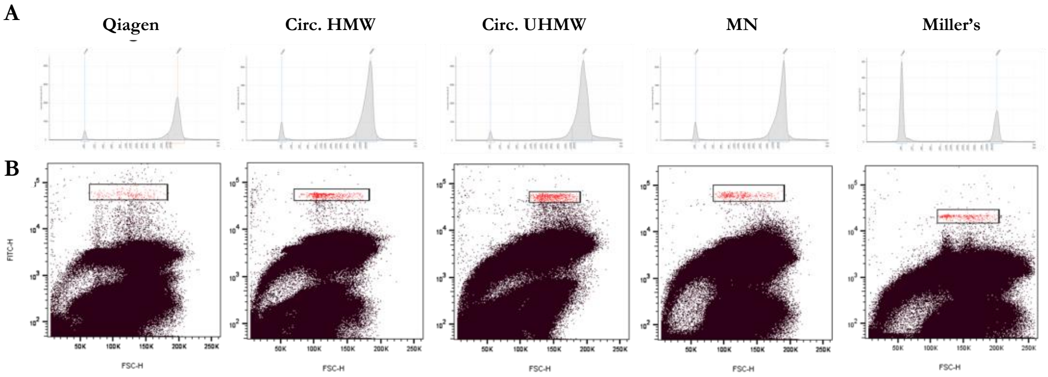


Figure 10: Comparison of DNA fragment size and flow-sorting from genomic DNA extracted with different methods. Genomic DNA was extracted with the Genomic Tip kit (Qiagen), Circulomics Nanobind CBB Big DNA Kit using either the HMW protocol (Circ. HMW) or the Ultra-HMW protocol (Circ. UHMW), the NucleoSpin Blood kit (Macherey-Nagel, MN), or Miller's protocol (Coriell samples). **(A)** Fragment distribution of starting genomic DNA samples obtained by capillary electrophoresis. **(B)** FACS dot plots showing forward scatter (FSC-H) *vs* fluorescence intensity (FITC-H) of droplets obtained after the dPCR step. The gate (red events) identifies the positive droplets that are sorted.

4.1.2 Illumina and ONT sequencing of Xdrop-enriched samples

Considering the high amount of recovered DNA (1.1 μg on average), both ONT and Illumina sequencing could be performed on a subset of samples representing three expansion states in *FMR1* microsatellite (healthy, pre-mutation and full mutation, **Table 6**). Sequencing generated on average 341,065 and 21,757,081 reads, with average lengths of 4,098 and 150 bp for ONT and Illumina, respectively (**Table 6**). However, a consistent fraction of reads was chimeric (59.3%) possibly derived from dMDA amplification as previously described[148,149]. Upon removing chimeric reads, primary alignment length was reduced to 1,506 bp (**Table 6** and **Figure 11B**, **C**). Since supplementary alignments belonged to the same genomic region, both alignments were considered in order to exploit the full dataset. The presence on chimeric reads limits the ability to accurately assess repeat lengths expanding over primary alignment sizes and further supports the unsuitability of the Xdrop workflow for the characterization of *DMPK* and *CNBP* microsatellites that extend far beyond the length of primary alignment. Low genome-wide coverage was achieved by both sequencing methods (0.16x and 0.7x on average) and significant enrichment was observed for all samples on the FMR1 gene: 357-fold for ONT and 467-fold for Illumina, on average (**Figure 11A, D** and **Table 6**). Importantly, reproducible enrichment was observed only on the FMR1 gene (**Figure 12A, B)**, supporting the specificity of the method. Maximum enrichment for both sequencing technologies was observed on the DS, and progressively decreased moving away from the target site, with a coverage > 10x maintained for up to ± 100 kbp flanking the DS (**Figure 11E, F**).

Average coverage on the gene body was 57x for ONT sequencing (**Table 6**), i.e. lower than the minimum threshold required to accurately call SNV using this technology[153]. On the contrary, Illumina sequencing achieved a significantly higher enrichment on the whole gene (362x), suggesting the Xdrop's potential to be used with Illumina also for SNVs and Indel characterization in *FMR1* gene body. More experiments would be required in order to confirm the accuracy of variant calling from dMDA-enriched samples, which however was beyond the scope of the present work. Further analyses were instead focused on the characterization of *FMR1* repeat expansions following ONT sequencing.

Table 6. (A) Enrichment and **(B)** sequencing statistics of Xdrop-enriched samples

A

Sample ID	Condition	Replicate	DNA extraction	Sorted droplets	Recovered DNA (ng)	Fold enrichment (qPCR)
NA12878	Normal	R1	Miller's	688	686	134
		R2		390	610	183
NA06891	Pre-mutation	R1	Miller's	800	1547	148
		R2		401	1738	300
NA20241	Pre-mutation	R1	Miller's	540	804	330
NA07537	Mutation	R1	Miller's	1135	1518	172
		R2		964	1130	601
NA12878	Normal	R1	Miller's	698	759	210
		R2		698	633	71
		R3		390	975	188
		R4		420	442.8	126.7
		R5		640	649	122.3
NA06891	Pre-mutation	R1	Miller's	317	530.4	508
		R2		243	423.8	228
NA20241	Pre-mutation	R1	Miller's	448	427	114
		R2		394	713	358
NA07537	Mutation	R1	Miller's	556	763	345
		R2		616	607	183

B

Sample ID	Replicate	Platform	# reads	Avg. cov. whole genome	# of reads *FMR1* gene	Avg. cov. *FMR1* gene	Fold enrichment (seq)	Avg. read length (bp)	Avg. primary alignment length (bp)
NA12878	R1	Nanopore	71,788	0.04x	1,062	12.17x	304	3,710	1,556
	R2		121,310	0.05x	1,644	17.66x	353	3,128	1,273
NA06891	R1	Nanopore	103,498	0.07x	1,510	15.04x	215	4,798	1,406
	R2		427,614	0.14x	5,722	66.05x	472	3,613	1,148
NA20241	R1	Nanopore	188,524	0.14x	5,188	75.34x	538	6,993	2,132
NA07537	R1	Nanopore	274,862	0.13x	3,302	36.26x	279	3,391	1,428
	R2		1,199,856	0.52x	15,534	175.24x	337	3,054	1,600
NA12878	R1	Illumina	19,053,800	0.68x	82,444	293.3x	431	150	135
	R2		19,818,606	0.67x	69,506	244.18x	360	150	134
	R3		44,456,772	1.37x	197,840	706.57x	515	150	133
	R4		3,116,556	0.10x	16,994	52.42x	372	150	115
	R5		5,088,850	0.17x	36,150	114.4x	448	150	119
NA06891	R1	Illumina	18,168,888	0.55x	73,754	257.3x	397	150	133
	R2		27,391,144	0.64x	48,572	164.1x	191	150	125
NA20241	R1	Illumina	22,675,960	0.78x	712,152	544.43x	698	150	128
	R2		31,580,602	1.17x	1,077,064	873.09x	746	150	132
NA07537	R1	Illumina	23,151,206	0.85x	421,844	336.4x	396	150	122
	R2		24,825,512	0.68x	514,456	399.94x	588	150	124

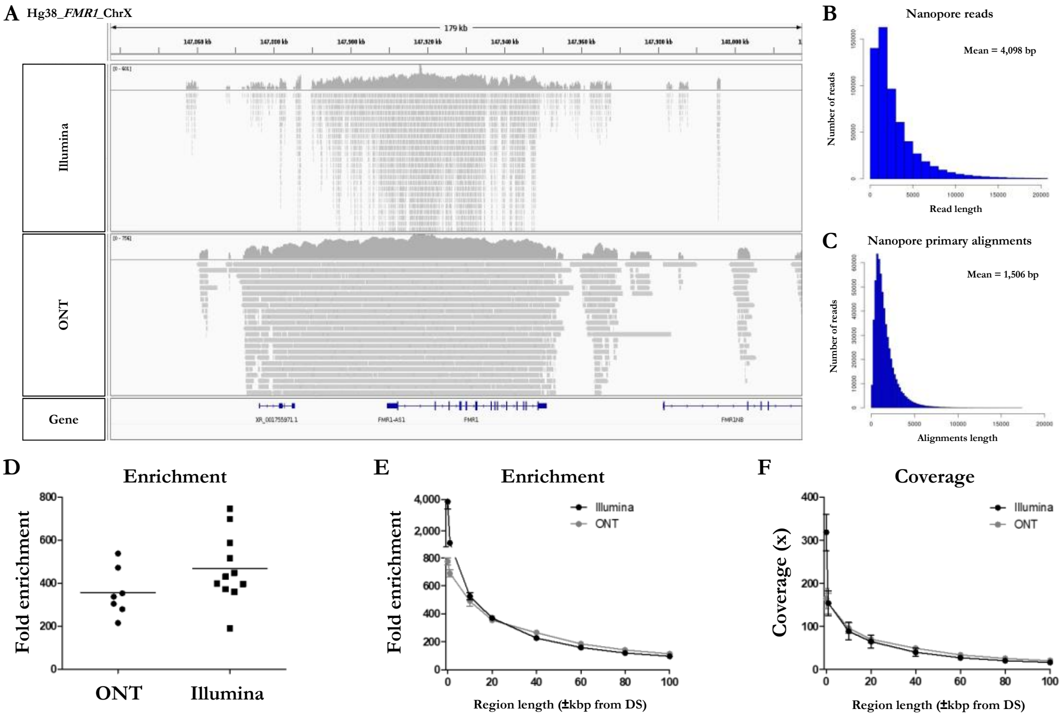


Figure 11. Statistics of FMR1 enrichment using the Xdrop technology. (A) Integrative Genomics Viewer (IGV) visualization of Illumina and ONT mapped reads obtained from a representative Xdrop-enriched sample. **(B)** Nanopore read and **(C)** primary alignment length distribution. **(D)** Comparison of fold enrichment on the whole *FMR1* gene between samples sequenced using Nanopore and Illumina platforms. **(E)** Fold enrichment and **(F)** average coverage following Nanopore and Illumina sequencing the Xdrop-enriched samples. Average coverage and Fold enrichment were calculated on the detection sequence (DS) and progressively up to 100 kbp upstream/downstream the DS.

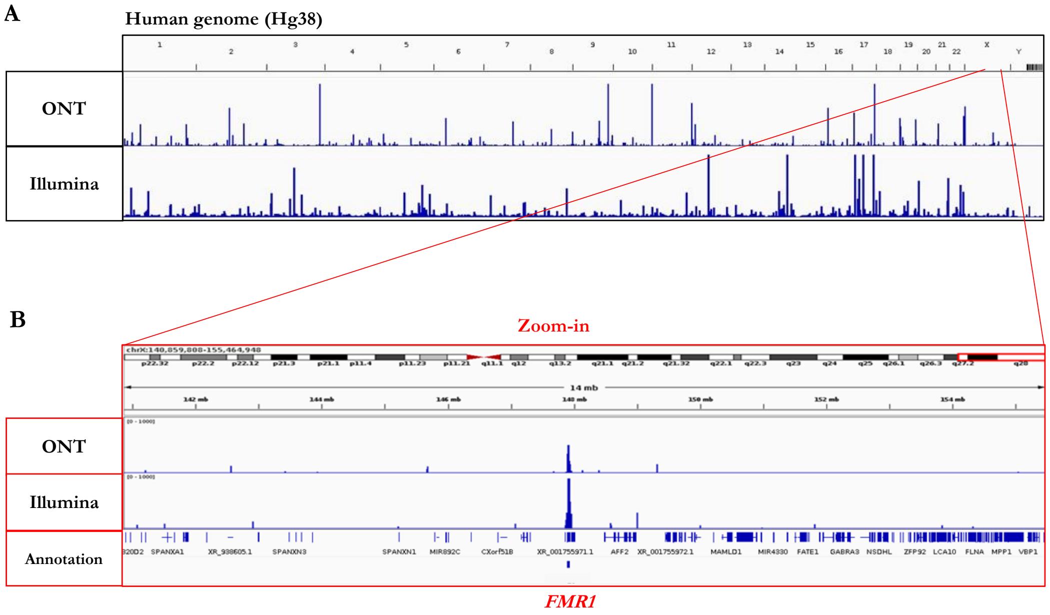


Figure 12. Coverage picture of Xdrop-enriched samples coupled to Nanopore and Illumina sequencing. (A) Integrative Genomics Viewer (IGV) visualization of a Xdrop-enriched representative sample following Nanopore and Illumina sequencing. **(B)** Zoom in showing reproducible enrichment only on the *FMR1* gene.

4.1.3 Characterization of FMR1 repeats by Xdrop enrichment and ONT sequencing

ONT sequencing data were analyzed from samples with known repeat features and showing expansions of 100–1000 bp (**Table 7**).

Table 7. Characterization of *FMR1* repeats from Xdrop-enriched samples by ONT sequencing. (A) ONT sequencing statistics and **(B)** comparison with previous reports. For each sample, Nanopore sequencing statistics covering the *FMR1* repeat are shown, along with the anticipated repeat features based on earlier reports and data generated from Xdrop-enriched samples.

A

ID	Condition	Sex	Total ONT reads	Mean coverage on repeat	Fold-enrichment on repeat	Reads spanning repeat
NA12878	Normal	F	96,549	33.0x	713.9	22
NA06891	Pre-mutation	M	265,556	55.5x	791.2	36
NA20241	Pre-mutation	F	188,524	330.8x	930.3	257
NA07537	Mutation	F	737,359	146.8x	912.9	87

B

ID	Allele	Average number of expected repeats	Average number of observed repeats
NA12878	**1**	28 CGG + 2 AGG	28 CGG + 2 AGG
	2	28 CGG + 2 AGG	28 CGG + 2 AGG
NA06891	**1**	118 (Wilson et al. 2008) 121 (Lim et al. 2017)	119 CGG + 1 AGG
	2	-	-
NA20241	**1**	29 (Juusola et al. 201; Chen et al. 2010; Lim et al. 2017) 27 CGG + 2AGG (Tsai et al.2017)	27 CGG + 2 AGG
	2	93-110 (Amos Wilson et al. 2008) 125 (Lim et al. 2017) 119, mosaicism (Tsai et al. 2017)	114 CGG + 1 AGG (mosaicism)
NA07537	**1**	28-29 (Adler et al. 2011) 27 CGG + 2 AGG (Tsai et al. 2017)	27 CGG +2 AGG
	2	>200, mosaicism (Adler et al. 2011; Lim et al. 2017; Tsai et al. 2017)	342 CGG + 1 AGG (mosaicism)

The consistent enrichment achieved on the target (range 215 – 538x) facilitated the extraction of sufficient reads spanning the entire tandem array (22 to 257) and allowed us to determine allele counts and features for every sample (**Figure 13A, B** and **Table 7)**. Sample NA12878 showed the anticipated normal pattern of 28 CGG repeats in both alleles, interrupted by the AGG trinucleotide at two sites. Sample NA06891 was derived from a male patient in the pre-mutation stage, with 118–121 CGG repeats according to previous sequencing data[154,155]. Consistently, our analysis counted an average of 119 CGG repeats and highlighted the presence of a single AGG trinucleotide interrupting the array. Sample NA20241 was obtained from a female patient heterozygous for normal and pre-mutated alleles. The expanded allele was reported to contain 93–110 repeats based on traditional methods[154] whereas more recent PacBio sequencing analysis revealed two groups of molecules with 90 and 120 repeats, respectively[43]. In agreement with the latter study, our analysis demonstrated the presence of mosaicism in this sample, evident as a bimodal distribution of sequencing read lengths, with modal values of 92 and 113 repeats. The CGG repeat count of the normal allele was also confirmed as 29, interrupted by two AGG trinucleotides. Sample NA07537 was previously reported to be heterozygous with 29 CGG repeats in the normal allele and > 200 in the expanded allele, corresponding to a full mutation[134]. The expanded allele was also characterized by PacBio sequencing, revealing a broad size distribution of 272–400 CGG repeats, which was confirmed by our data. Specifically, ONT sequencing reads ranged from a minimum of 196 to a maximum of 402 repeats, with a modal value of 342.

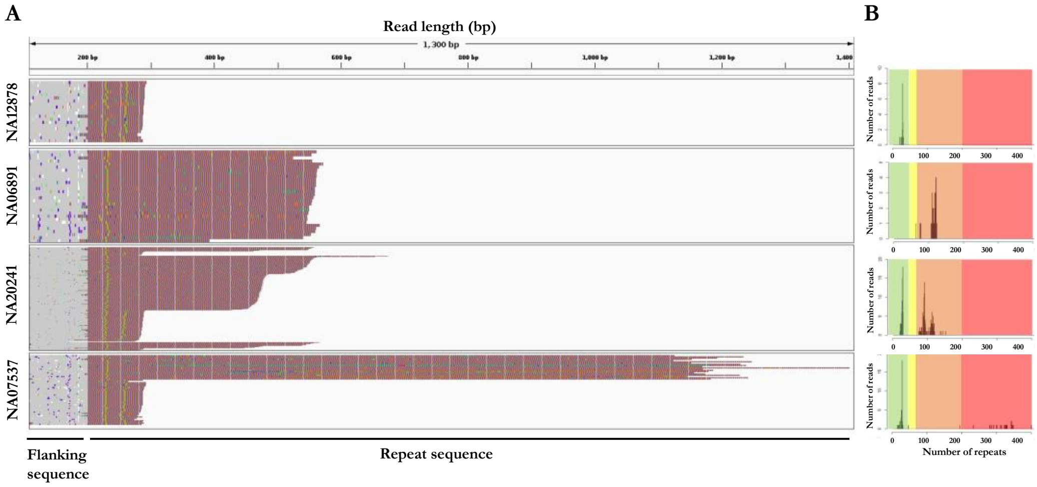


Figure 13. Visualization of repeat structure and length after sequencing Xdrop-enriched samples on the ONT platform. (A) Individual ONT reads were trimmed to include only the *FMR1* repeat region plus 400bp flanking sequence and aligned at the repeat 5'-end. Each line represents a single read, colored according to: A=green, T=red, G=orange, and C=blue. **(B)** Repeat count histograms showing the number of reads reporting a certain repeat length: shaded background in each plot represents risk ranges for disease development. Green=normal; yellow=intermediate; orange=pre-mutation; and red=full mutation.

4.2 REPEAT CHARACTERIZATION BY CAS9-MEDIATED ENRICHMENT

Limitations of the Xdrop method were mostly related to the dMDA step, that was likely responsible of shortening the recovered target DNA after amplification. Hence, in order to fully characterize *DMPK* and *CNBP* larger expansions, we tested the Cas9-mediated capture. The method is indeed an amplification-free assay and, being coupled with ONT long-read sequencing, can potentially generate very long reads (up to 2.3 Mb[9]).

4.2.1 DMPK microsatellite enrichment via Cas9-mediated capture and ONT sequencing

Assay validation was first performed on the *DMPK* locus. A set of 5 gRNAs were designed on both the flanking regions of the *DMPK* microsatellite (**Figure 14A**). The cut efficiency of each gRNA was first validated by qPCR using a method that was developed and tested for the first time in this thesis work. As described more in details in the Material and Method section, primer pairs were designed on the flanking region of each cut site and the accumulation of PCR product was monitored by qPCR in comparison to an uncut sample (**Figure 14B**). According to this approach, we demonstrated that all gRNAs tested, but one, had cut efficiency > 60% (**Figure 14C** and **Table 5.** The four most efficient gRNAs (1, 2, 3 and 4) were tested further via ONT sequencing. The gRNAs were paired to make sure that the enriched fragment length was >3 kbp because during library preparation fragments <3 kbp are removed. Hence, gRNA 1 was paired with gRNA 4 to enrich a fragment of 6.5 kbp whereas gRNA 2 was paired to gRNA 3 to enrich a fragment of 9.6 kbp. To prevent any cut interference between gRNAs on the same end of the target, cut reactions were performed in parallel and then pooled prior to ONT sequencing. The assay was performed on gDNA extracted from the HEK293 cell line, known to carry not-expanded alleles (http://hek293genome.org/v2/)[156].

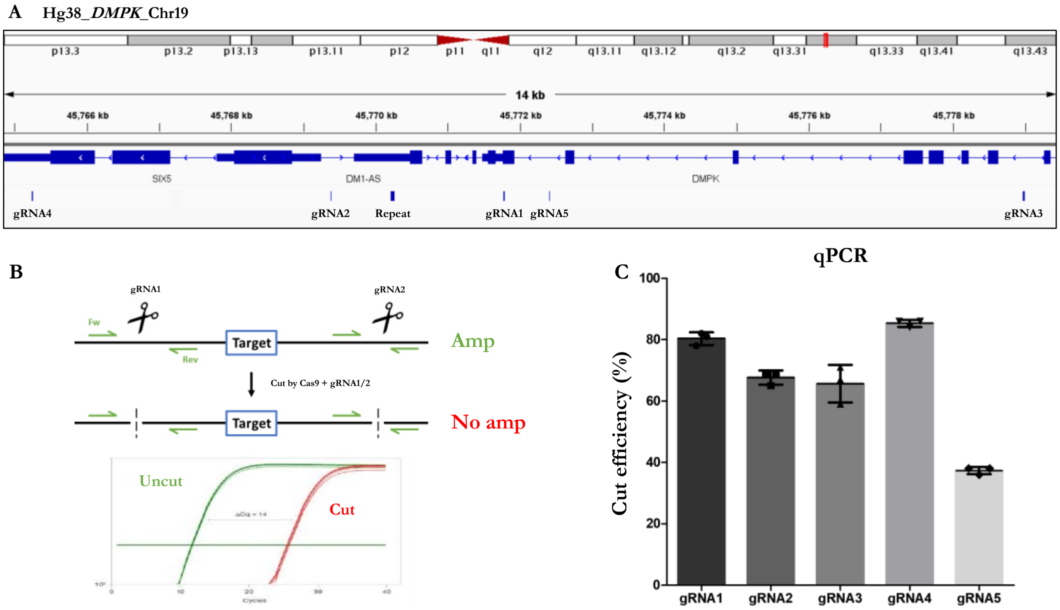


Figure 14. gRNA design and cut efficiency evaluation via qPCR. (A) Integrative Genomics Viewer (IGV) visualization showing the position of the *DMPK* repeat and the gRNAs on the flanking regions. **(B)** Schematic representation of the qPCR assay used to evaluate the cut efficiency of the gRNAs. **(B)** Primer pairs were designed flanking each gRNA cut site and then used to amplify intact or cut DNA, respectively. The amount of intact versus cut DNA is compared and plotted as cut efficiency (i.e. % of cut DNA). Prolactin (non-target gene) was used for DNA input normalization. **(C)** gRNA cut efficiency plotted as % of cut DNA. (N=3).

Sequencing data analysis showed that both gRNA pairs performed clear cuts on both sides of the target (**Figure 15A**). However, we observed a significant difference in the cut efficiency. In agreement with qPCR results, gRNA pair 1-4 produced significantly higher on-target reads (385, 0.59%) as compared to gRNA pair 2-3 (40, 0.06%) (**Figure 15B**). This resulted in a 1,833-fold enrichment for gRNA pair 1-4 and 190-fold for pair 2-3 (**Figure 15C**). Based on these data, we could determine that > 70% cut efficiency threshold indicated good-performing gRNAs. Therefore, gRNA pair 1-4 was selected for further validation and ONT sequencing.

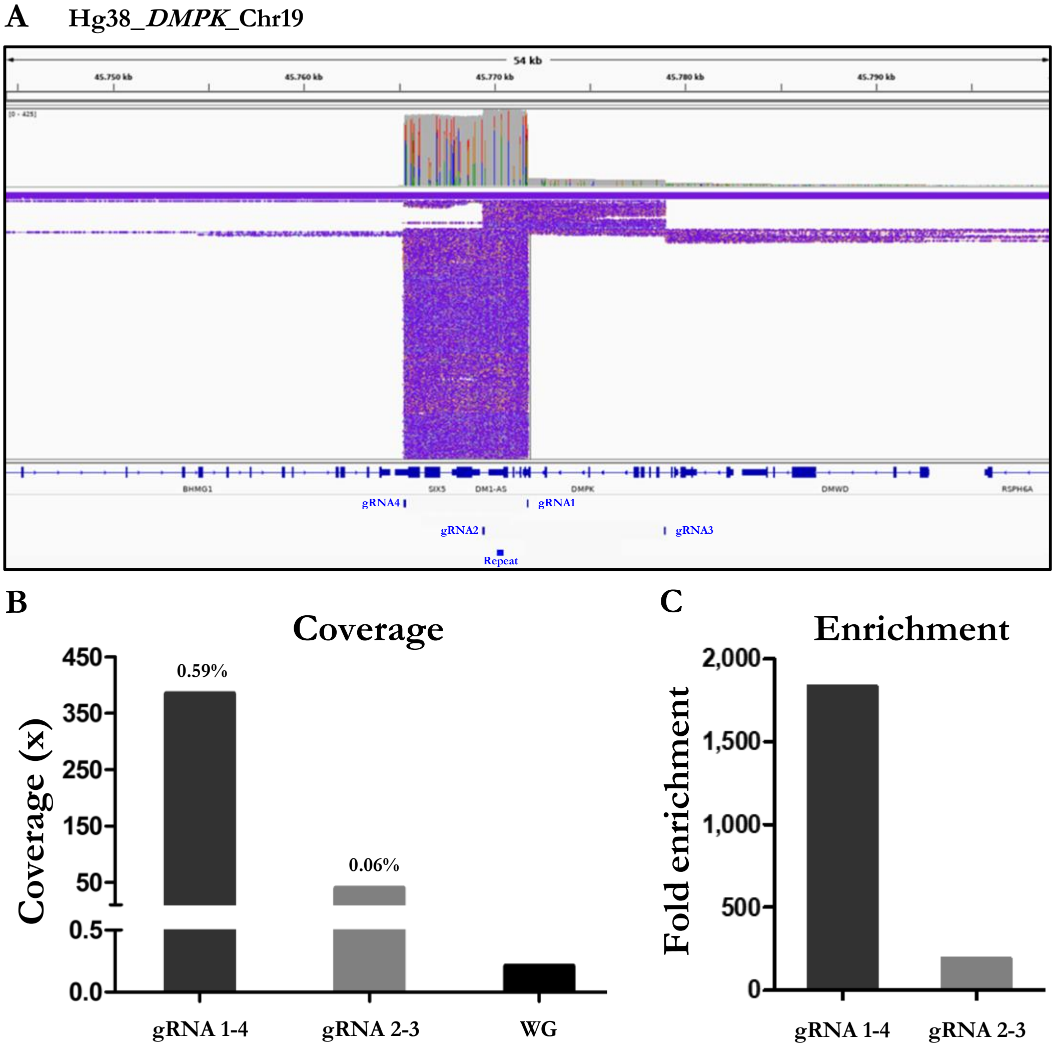


Figure 15. Cas9-mediated enrichment and Nanopore sequencing of the *DMPK* microsatellite from a healthy control. (A) Integrative Genomics Viewer (IGV) visualization of nanopore reads on the DMPK microsatellite. The position of the gRNAs and the microsatellite are indicated at the bottom. Both the enriched targets (gRNA1-4 and gRNA2-3) displays clear cuts on both sides, tough to different degrees. **(B)** Average coverage data for the *DMPK* microsatellite and whole genome (WG, i.e.

background coverage). Values on top indicate the % of on-target reads. **(C)** Fold enrichment of the *DMPK* microsatellite.

We then tested whether the HMW DNA extraction method could have some interference with gRNA cut efficiency. At this aim, the cut efficiency of gRNA pair 1-4 was tested through qPCR and across four different HMW gDNA extraction methods yielding different DNA fragment sizes (**Figure 16A**). No significant differences were shown between the methods tested in terms of cut efficiency, with the exception of Circulomics U-HMW protocol (**Figure 16B**). The entangled, highly viscous nature of this gDNA may have interfered with the cut reaction, resulting in a highly variable efficiency for gRNA 1. Based on these results, we showed that any HMW DNA extraction method was suitable for Cas9-mediated capture, as long as the average fragment length is higher than the enriched target length. In contrast Ultra-HMW protocol should be avoided as they could lead to poor cut-efficiency and thus enrichment variability.

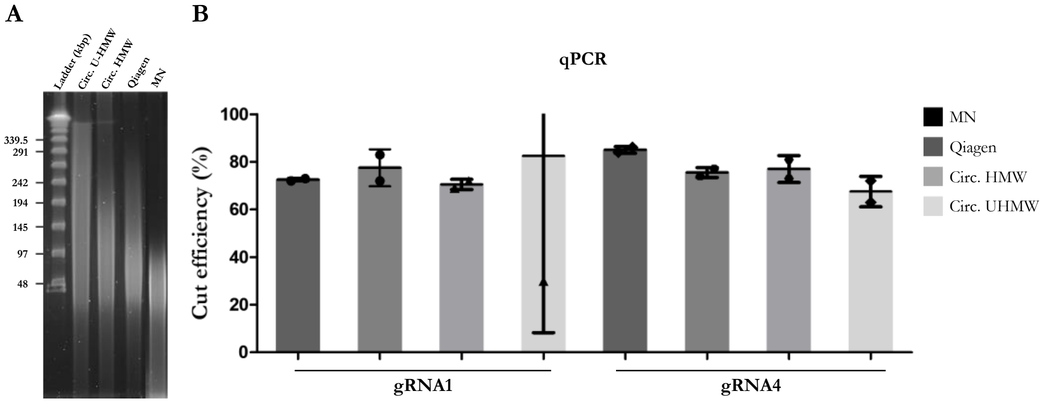


Figure 16. Cas9-mediated cut efficiency with different HMW DNA extraction methods. (A) Pulsed-field gel electrophoresis of genomic DNA extracted with Circulomics Nanobind CBB Big DNA Kit using either the HMW protocol (Circ. HMW) or the Ultra-HMW protocol (Circ. UHMW), the Genomic Tip kit (Qiagen), or the NucleoSpin Blood kit (Macherey-Nagel, MN). **(B)** Cut efficiency of gRNAs 1 and 4 tested across the 4 different DNA extraction methods (N=2), from standard to U-HMW. The amount of intact versus cut DNA is compared and plotted as cut efficiency (i.e. % of cut DNA). Prolactin (non-target gene) was used for DNA input normalization.

In order to validate the method for the characterization of *DMPK* microsatellite on patients with confirmed DM1 diagnosis, we extracted DNA from a set of four selected samples. However, no blood sample yielded good quality gDNA in terms of integrity

(Peak < 60 kbp, DIN < 7) and quantity (< 5 ug). This could be ascribed to inappropriate storage of blood samples (i.e. several freeze-thaw cycles) which inevitably led to DNA degradation. To confirm/reject this hypothesis, we decided to test the method on the best sample. However, ONT sequencing generated only 15 on-target reads, corresponding to a fold enrichment of 16.6-fold and confirming our hypothesis. Since the use of degraded DNA was highly risky and could have led to suboptimal results, we did not continue experiments on DM1 patients. Additional blood samples are therefore required to validate the Cas9-*DMPK* assay for the analysis of pathogenic repeats.

4.2.2 CNBP microsatellite enrichment by Cas9-mediated capture and ONT sequencing

Based on the same criteria used for *DMPK*, we designed gRNAs on the flanking sequences of the *CNBP* microsatellites (**Figure 17A**). The gRNAs displayed >70% cut efficiency as determined by qPCR (**Figure 17B** and **Table 5**), that according to previous tests on *DMPK* allowed to select gRNA with optimal efficiency for targeted ONT-sequencing. Confirmation of gRNA cut efficiency was performed by ONT sequencing on a healthy control (HEK293 cell line) and including gRNA pair 1-4 for *DMPK* as internal control. A total of 65,154 PASS reads were generated, of which 1,595 (2.45%) were on-target (**Table 8**). The gRNA pairs designed on *CNBP* generated clear cuts on both sides of the target (**Figure 17C**) and provided up to1,170 (1.8%) on-target reads (**Figure 17D**). Whole genome coverage was 0.21x, resulting in a 5,571-fold enrichment of the *CNBP* microsatellite (**Figure 17E, F**).

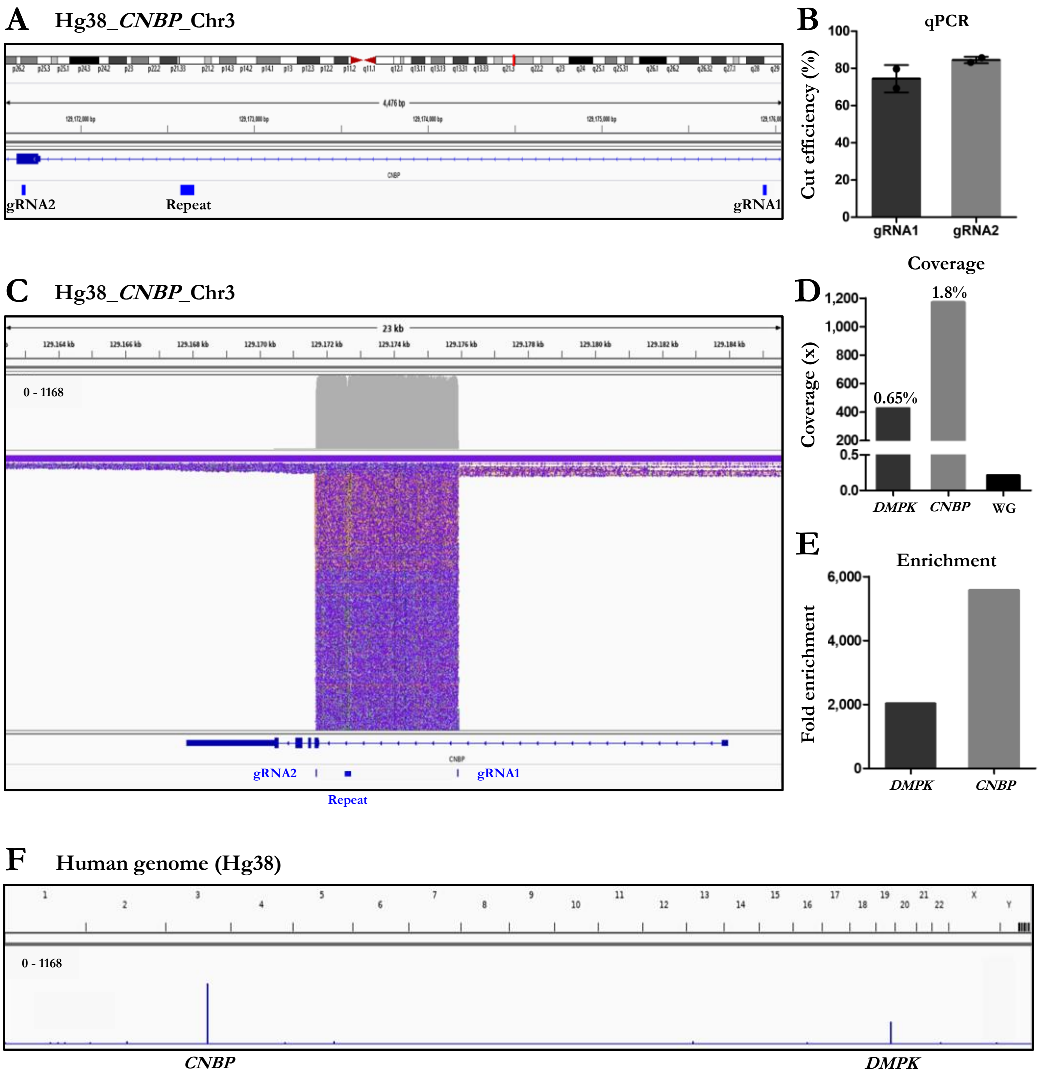


Figure 17. Cas9-mediated enrichment and Nanopore sequencing of the *CNBP* microsatellite from a healthy control. (A) Integrative Genomics Viewer (IGV) visualization showing the position of the *CNBP* repeat and the gRNAs on the flanking regions. **(B)** Cut efficiency of gRNA pairs 1-2 designed to excise the *CNBP* microsatellite. The gRNAs were pooled together and used to simultaneously excise the target from gDNA (N=2). The amount of intact versus cut DNA is compared and plotted as cut efficiency (i.e. % of cut DNA). Prolactin (non-target gene) was used for DNA input normalization. **(C)** Integrative Genomics Viewer (IGV) visualization of nanopore reads on the *CNBP* microsatellite. The position of the gRNAs and the repat are indicated at the bottom. The enriched target displays clear cuts on both sides, accounting for a good performance of Cas9-mediated capture. **(D)** Average coverage data at the *CNBP* microsatellite and at whole genome level (WG, i.e. background coverage). Values on top indicate the % of on-target reads. **(E)** Fold enrichment data of the *CNBP* microsatellite. For comparison, statistics on the DMPK microsatellite are also shown. **(F)** Integrative Genomics Viewer (IGV) visualization showing the genome-wide coverage following Cas9-mediated capture and Nanopore sequencing. Off-target reads are distributed randomly across the genome and result from the ligation of nanopore adapters to random breakage points.

Table 8. Enrichment and sequencing statistics of Cas9-enriched samples. Reported are the statistic of ONT sequencing runs from **(A)** HEK293 cell line, **(B)** DM2 patients in singleplex and **(C)** DM2 patients in multiplex.

A

		Healthy control		
Run ID		Exp0		
Source		HEK293 cell line		
# samples		1		
input		10 ug		
Total aligned PASS reads		65,154		
On-target reads		1,595		
On-target %		2.45%		
Whole genome avg. cov.		0.21x		
Target	**Target length (hg38)**	**Avg coverage**	**On-target reads**	**%**
DMPK	6.5 kbp	425x	425	0.65%
CNBP	4.2 kbp	1170x	1,170	1.8%

B

		DM2 (*CNBP*) patients - Singleplex runs											
Run ID		Exp1			Exp2			Exp3			Exp4		
Source		Whole blood			Whole blood			Whole blood			Whole blood		
# samples		1			1			1			1		
input		5 ug			2 ug			7 ug			5 ug		
Total aligned PASS reads		154,046			35,025			21,550			60,315		
On-target reads		705			359			414			146		
On-target %		0.46%			1.02%			1.92%			0.24%		
Whole genome avg. cov.		0.14x			0.077x			0.07x			0.07x		
Target	**Target length (hg38)**	**Avg coverage**	**On-target reads**	**%**	**Avg coverage**	**On-target reads**	**%**	**Avg coverage**	**On-target reads**	**%**	**Avg coverage**	**On-target reads**	**%**
DMPK	6.5 kbp	121.2x	122	0.1%	79x	79	0.2%	102x	102	0.5%	18.5x	19	0.03%
CNBP	4.2 kbp	584.2x	624	0.4%	279.8x	283	0.8%	311.9x	330	1.5%	127.6x	142	0.24%

C

		DM2 (*CNBP*) patients - Multiplex runs											
Run ID		Exp1			Exp2			Exp3			Exp4		
Source		Whole blood			Whole blood			Whole blood			Whole blood		
# samples		5			4			4			3		
input		2 ug / sample			1 - 4 ug / sample			3 - 5 ug / sample			4.5 - 10 ug / sample		
Total aligned PASS reads		1,284,125			819,341			98,083			552,544		
On-target reads		579			391			88			566		
On-target %		0.05%			0.05%			0.09%			0.10%		
Whole genome avg. cov.		0.94x			0.6x			0.118x			0.66x		
Target	**Target length (hg38)**	**Avg coverage**	**On-target reads**	**%**	**Avg coverage**	**On-target reads**	**%**	**Avg coverage**	**On-target reads**	**%**	**Avg coverage**	**On-target reads**	**%**
DMPK	6.5 kbp	71x	71	0.01%	33x	33	0.004%	10x	10	0.010%	71.23x	72	0.013%
CNBP	4.2 kbp	507.89x	547	0.04%	357.57x	397	0.048%	78.3x	84	0.086%	494.4x	501	0.091%

Based on these results, we moved on to apply the method on gDNA extracted from 9 patients with confirmed DM2 diagnosis, which have been characterized using traditional approaches by the Medical Genetics Section of Policlinico Tor Vergata (**Table 9**). Cas9-mediated capture and ONT sequencing were performed through a total of 4 single-plex and 4 multiplex runs (**Table 8**). The gRNA pair 1-4 for *DMPK* were always included as internal control. Enrichment performances on the *CNBP* microsatellite were lower than in the pilot experiment (HEK293 cell line), as previously reported for DNA extracted from clinical samples compared to cell lines[64], but still consistently high (average coverage: 343x) and associated to 2,079-fold enrichment on average for all experiments (**Table 8**). Multiplexing runs showed higher background and lower enrichment as compared to the singleplex experiments (**Figure 18A, B**).

Table 9. Repeat expansion analysis of DM2 patients using traditional diagnostic approaches. Clinical features are presented along with the anticipated repeat features based on standard diagnostic procedures (Sanger sequencing & Southern blotting). Blood samples, Sanger consensus sequences and Southern blot-derived expansion lengths were provided by the University of Tor Vergata**.**

				Normal Allele (Sanger)		Expanded allele (Southern blot)	
Sample ID	**Sex**	**Age**	**Age at onset**	**Repeat lenght (bp)**	**Repeat structure**	**Approximate repeat length (bp)** *(min - max)*	**Repeat structure**
DM2.A1	F	75	70	136	(TG)24 (TCTG)7 (CCTG)5 GCTG CCTG TCTG (CCTG)7	32,500 - 40,000	--
DM2.A2	M	27	25	130	(TG)17 (TCTG)9 (CCTG)5 GCTG CCTG TCTG (CCTG)7	14,183 - 20,000	--
DM2.A3	M	21	-	132	(TG)20 (TCTG)8 (CCTG)5 GCTG CCTG TCTG (CCTG)7	14,183 - 20,323	--
DM2.A4	M	65	61	122	(TG)19 (TCTG)9 (CCTG)12	29,027 - 32,745	--
DM2.B	F	49	44	134	(TG)21 (TCTG)7 (CCTG)6 GCTG CCTG TCTG (CCTG)7	31,000 - 40,000	--
DM2.C	M	20	-	140	(TG)24 (TCTG)6 (CCTG)7 GCTG CCTG TCTG (CCTG)7	22,000 - 29,027	--
DM2.D	M	44	39	134	(TG)19 (TCTG)9 (CCTG)5 GCTG CCTG TCTG (CCTG)7	18,000 - 20,000	--
DM2.E	F	61	43	134	(TG)19 (TCTG)9 (CCTG)5 GCTG CCTG TCTG (CCTG)7	25,000 - 30,000	--
DM2.F	M	56	50	138	(TG)21 (TCTG)9 (CCTG)5 GCTG CCTG TCTG (CCTG)7	20,323 - 39,000	--

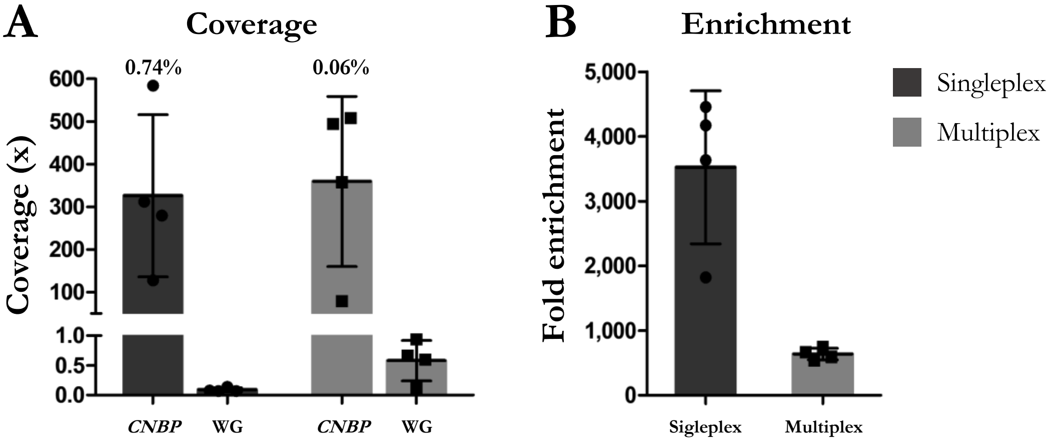


Figure 18. Cas9-mediated enrichment coupled to Nanopore sequencing of the *CNBP* microsatellite from DM2 patients. (A) Average coverage and **(B)** Fold-enrichment data on the *CNBP* microsatellites from singleplex (N = 4) and multiplex (N = 4) nanopore runs. Values on top indicate the % of on-target reads.

Considering average values across all samples, a total 105,737 PASS reads were generated, of which 308 (0.3%) were on-target (**Table 10**). Among these, 186 reads were fully spanning either the normal (N=145, 78%) or the expanded (N=41, 22%) allele (**Figure 19A** and **Table 10**). Reads derived from the normal allele were *de novo* assembled to obtain an accurate consensus sequence. The complex $(TG)_v(TCTG)_w(CCTG)_x(NCTG)_y(CCTG)_z$ array was correctly identified in all patients and ranged from 122 to 141 bp, corresponding to 12 – 15 CCTG tetraplets (**Figure 19B** and **Table 10**).

Table 10. Nanopore sequencing statistics and *CNBP* repeat analysis. For each sample, Nanopore sequencing statistics covering the *CNBP* repeat are shown, along with normal and expanded allele features derived from data analysis. The last columns shows the % of reads from the expanded allele carrying the TCTG "atypical" motif.

A

			On-target average coverage			
			Total	Spanning whole repeat		
Sample ID	Total PASS reads	On-target PASS reads		All	Ref	Alt
DM2.A1	149,569	216	197.7x	144x	113x	31x
DM2.A2	154,046	624	584.2x	354x	247x	107x
DM2.A3	197,651	105	98.9x	71x	60x	11x
DM2.A4	100,747	204	182.1x	113x	82x	31x
DM2.B	142,659	322	304.5x	137x	111x	26x
DM2.C	92,411	311	275.9x	189x	120x	69x
DM2.D	37,975	176	171.8x	120x	93x	27x
DM2.E	35,025	283	279.8x	185x	153x	32x
DM2.F	41,552	528	507.7x	365x	330x	35x

B

	Normal allele		Expandedallele		
Sample ID	Repeat length (bp)	Repeat structure	Repeat length (bp) *(min - max)*	Repeat structure	# reads carrying TCTG
DM2.A1	136	(TG)24 (TCTG)7 (CCTG)5 GCTG CCTG TCTG (CCTG)7	3,241 - 46,685	(TG)20 (TCTG)7 (CCTG)1,000-12,000 (TCTG)0-10	14 (45%)
DM2.A2	131	(TG)17 TGCTG (TCTG)8 (CCTG)5 GCTG CCTG TCTG (CCTG)7	864 - 23,779	(TG)18 (TCTG)7 (CCTG)1,000-4,500 (TCTG)0-2,000	92 (86%)
DM2.A3	132	(TG)20 (TCTG)8 (CCTG)5 GCTG CCTG TCTG (CCTG)7	4,429 - 18,983	(TG)19 (TCTG)7 (CCTG)3,000-5,000 (TCTG)0-25	8 (73%)
DM2.A4	122	(TG)19 (TCTG)9 (CCTG)12	660 - 34,284	(TG)18 (TCTG)7 (CCTG)250-8,000 (TCTG)0-1,500	9 (29%)
DM2.B	134	(TG)21 (TCTG)7 (CCTG)6 GCTG CCTG TCTG (CCTG)7	344 - 23,358	(TG)18 (TCTG)7 (CCTG)300-4,000 (TCTG)0-400	6 (23%)
DM2.C	141	(TG)24 TGCTG (TCTG)5 (CCTG)7 GCTG CCTG TCTG (CCTG)7	700 - 31,753	(TG)20 (TCTG)7 (CCTG)150-8,000	0 (0%)
DM2.D	138	(TG)21 (TCTG)9 (CCTG)5 GCTG CCTG TCTG (CCTG)7	383 - 19,143	(TG)18 (TCTG)7 (CCTG)100-4,000 (TCTG)0-1,000	3 (11%)
DM2.E	134	(TG)19 (TCTG)9 (CCTG)5 GCTG CCTG TCTG (CCTG)7	848 - 25,162	(TG)18 (TCTG)6 (CCTG)200-6,200	0 (0%)
DM2.F	138	(TG)21 (TCTG)9 (CCTG)5 GCTG CCTG TCTG (CCTG)7	1,533 - 32,824	(TG)15 (TCTG)10 (CCTG)400-6,000 (TCTG)0-2,000	15 (43%)

Both the size and the repeat pattern identified in each patient were in large agreement with results from Sanger sequencing (99.5% mean accuracy, Pearson's r = 0.971, P-value <0.0001), with 6 patients showing a perfect match, 2 displaying a single-nucleotide difference and only one a di-nucleotide difference (**Figure 19D**, **Table 10 - 11**).

Table 11. Comparison between Sanger and ONT consensus sequences. Differences are highlighted in red. The last columns shows the identity % between the consensus sequences.

	Sanger consensus		Nanopore consensus		
Sample ID	**Repeat lenght (bp)**	**Repeat structure**	**Repeat length (bp)**	**Repeat structure**	**Identity with Sanger**
DM2.A1	136	(TG)24 (TCTG)7 (CCTG)5 GCTG CCTG TCTG (CCTG)7	136	(TG)24 (TCTG)7 (CCTG)5 GCTG CCTG TCTG (CCTG)7	100%
DM2.A2	130	(TG)17 (TCTG)9 (CCTG)5 GCTG CCTG TCTG (CCTG)7	131	(TG)17 TGCTG (TCTG)8 (CCTG)5 GCTG CCTG TCTG (CCTG)7	99.2%
DM2.A3	132	(TG)20 (TCTG)8 (CCTG)5 GCTG CCTG TCTG (CCTG)7	132	(TG)20 (TCTG)8 (CCTG)5 GCTG CCTG TCTG (CCTG)7	100%
DM2.A4	122	(TG)19 (TCTG)9 (CCTG)12	122	(TG)19 (TCTG)9 (CCTG)12	100%
DM2.B	134	(TG)21 (TCTG)7 (CCTG)6 GCTG CCTG TCTG (CCTG)7	134	(TG)21 (TCTG)7 (CCTG)6 GCTG CCTG TCTG (CCTG)7	100%
DM2.C	140	(TG)24 (TCTG)6 (CCTG)7 GCTG CCTG TCTG (CCTG)7	141	(TG)24 TGCTG (TCTG)5 (CCTG)7 GCTG CCTG TCTG (CCTG)7	99.3%
DM2.D	134	(TG)19 (TCTG)9 (CCTG)5 GCTG CCTG TCTG (CCTG)7	138	(TG)21 (TCTG)9 (CCTG)5 GCTG CCTG TCTG (CCTG)7	97.1%
DM2.E	134	(TG)19 (TCTG)9 (CCTG)5 GCTG CCTG TCTG (CCTG)7	134	(TG)19 (TCTG)9 (CCTG)5 GCTG CCTG TCTG (CCTG)7	100%
DM2.F	138	(TG)21 (TCTG)9 (CCTG)5 GCTG CCTG TCTG (CCTG)7	138	(TG)21 (TCTG)9 (CCTG)5 GCTG CCTG TCTG (CCTG)7	100%

Reads derived from the expanded alleles showed wide length variability, ranging from 344 bp up to as much as 46.6 kbp (**Figure 19C** and **Table 10**), confirming the presence of extremely large expansions in these patients. To our knowledge, the latter represents the longest repeat expansion analyzed so far at the single-nucleotide resolution[45–48] and one of the longest DNA fragment captured using the Cas9-mediated enrichment with no specific adjustment [64,68] .

Considering average values per sample, the number of repetitive tetraplets varied from 1,371 to 4,421, which corresponded to expansion lengths ranging from 5,485 bp up to 17,685 bp. Moreover, the length of the longest expanded molecule sequenced with ONT was in large agreement with data derived from Southern blotting analysis (Pearson's r = 0.6840, P-value = 0.0422, **Figure 19E, Table 9-10**), with the exception of 1 sample (DM2.B). Read sizes derived from the expanded allele were very variable even within the same individual (intra-donor variability of 65% on average, **Figure 19C**), suggesting the presence of a pronounced mosaicism in agreement with previous reports in DM2[115,122,123].

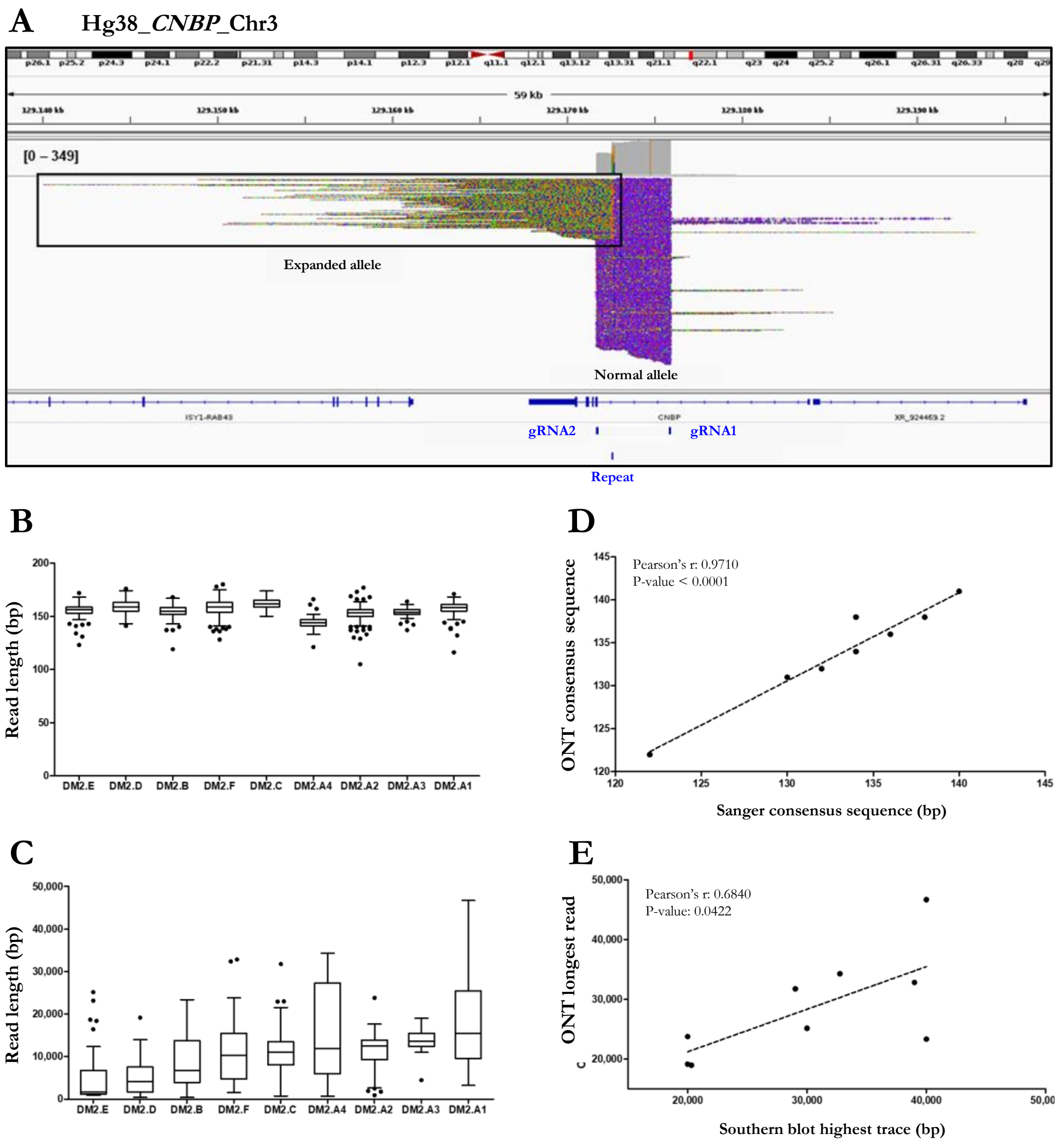


Figure 19. *CNBP* repeat analysis of DM2 patients. (A) Integrative Genomics Viewer (IGV) visualization of the enriched *CNBP* on Chromosome 3 from a representative DM2 sample (DM2.A2). The normal allele displays clear cuts on both sides while the expanded allele is represented by the soft-clipped reads not matching to the reference. **(B)** Read length distributions of the normal alleles. **(C)** Read length distributions of the expanded alleles. The distributions support the expansion mosaicism previously reported in literature. Box = Interquartile range (IQR), horizontal line = median, whiskers = upper quartile +1.5 x IQR and lower quartile – 1.5 x IQR, dots = outliers. **(D)** Correlations between ONT consensus and Sanger consensus sequences from the normal allele (Person's r = 0.971, P-value <0.0001, N = 9). **(E)** Correlations between ONT sequencing and Southern blot data from the expanded allele (Person's r = 0.7604, P-value = 0.0174, N = 9). The plot has been generated by considering the longest complete read from ONT data and the upper edge of the Southern blot trace.

To identify the repeat pattern characterizing the expanded microsatellite locus, the occurrence of tetra-nucleotides motives was therefore identified in each individual read, and highlighted using distinct colors, after aligning "complete sequences" at the 5' and 3'- end (**Figure 20A-B**). The uninterrupted $(TG)_v(TCTG)_w(CCTG)_x$ motif, known to characterize expanded *CNBP* alleles, was recognized at the 5' of the repeat locus in all patients (**Figure 20A-B** and **Table 10**). However, only 2 patients contained this "pure" pattern of $(CCTG)_n$ repetitions; the remaining 7 carried also an additional $(TCTG)_n$ repeated array (colored in red) recurrent at the 3'-end of the CCTG expansion, which has been never reported in DM2 patients previously (**Figure 20A-B** and **Table 10**). When present, the TCTG expanded motif was detected in a very variable fraction of sequences (11% to 86% of expanded allele reads, **Figure 20C** and **Table 10**), and presented highly diverse length, both between donors and within the same sample, ranging from 40 bp to 8,000 bp (**Figure 20D** and **Table 10**).

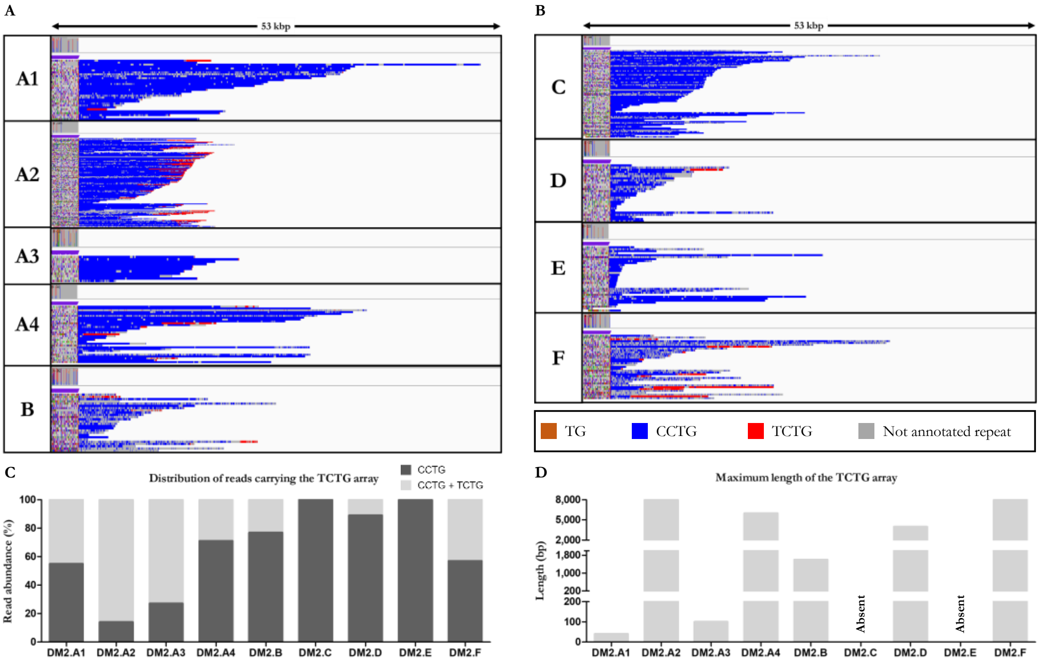


Figure 20. In depth analysis *CNBP* expanded alleles. (A-B) Integrative Genomics Viewer (IGV) visualization of the expanded alleles. Complete reads are aligned at the 5'-end (figure) and subsequently at the 3'-end (not shown). The visualization window is set at 53 kbp. Each repeated motif is substituted by a stretch of a single nucleotide, which was then assigned a color to better discriminate the repeat (see repeat color code). Only samples DM2.C e DM2.E display the "pure" CCTG (blue) expansion whereas all the others display the TCTG motif (red) right downstream the CCTG one. **(C)** Distribution of reads carrying the TCTG array. **(D)** Maximum length of the TCTG array. The TCTG motif displayed within-sample length variability and did not occur in all the sequences from the same sample.

4.3 REPEAT CHARACTERIZATION BY HYBRIDIZATION CAPTURE APPROACHES

The last long-DNA target enrichment approach was tested only on the *DMPK* microsatellite, that was not successfully characterized with the previous two approaches tested. In particular, we tested the performances of three different hybridization-based methods exploiting biotinylated **a)** PNA, **b)** dCas9 and **c)** dsDNA-probes. A protocol suitable for all methods was developed, whose main steps were **I)** Probe-DNA hybridization, **II)** Target capture, **III)** Off-target removal, **IV)** Probe-target DNA separation and **V)** Target DNA recovery. Each procedure is described in details in the M&M section. Each method was tested on gDNA fragmented at 5 kbp and extracted from a healthy control (HEK293 cell line) using the Circulomics HMW method. Enrichment performances were validated via qPCR and using a primer pair annealing ~700 bp upstream the *DMPK* microsatellite (**Figure 21**).

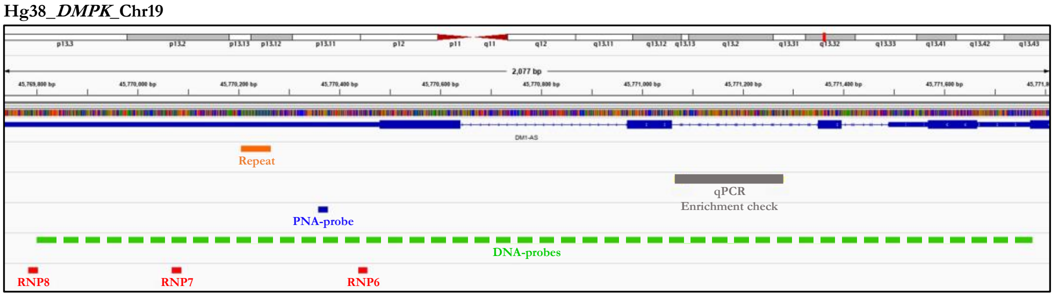


Figure 21. Hybridization-based capture approaches. Integrative Genomics Viewer (IGV) visualization of the three hybridization-based capture approaches reported in this thesis: PNA-probes (blue), DNA-probes (green) and RNP-probes (i.e. dCas9 + gRNAs 6, 7 and 8, red). The figure also shows the localization of the *DMPK* microsatellite (orange) and of the primers used to assess the enrichment by qPCR (grey).

4.3.1 PNA-based enrichment

PNA-based enrichment involved the design of a biotinylated PNA-probe which annealed 93 bp upstream the *DMPK* microsatellite (**Figure 21**). Based on qPCR analysis, average on-target enrichment was 12-fold (**Figure 22B**) and recovered DNA was 5 ng (**Figure 22A**). In order to improve the recovery of target DNA for ONT sequencing, we tested different post-capture wash buffers with decreasing stringency (**Table 12**). As expected, the recovered target DNA increased as the buffer stringency decreased, yielding up to 15 ng when the washing step was omitted (W4) (**Figure 22A**). In turn, no significant improvement in the enrichment was reported (**Figure 22B**), with respect to standard conditions.

Table 12. Wash buffers used for the standard and optimized PNA-mediate capture protocols. For each buffer the table shows the relative stringency, the number of washes performed, the buffer composition, the salt composition, the presence/absence of EDTA and the pH.

Buffer	Stringency	# Wash	Composition	Salt	EDTA	pH
W1	• • • • •	I	25 mM Tris-HCl	150 mM NaCl	/	7.2
W1	• • • •	II	/	/	/	/
W3		I	25 mM Tris-HCl	150 mM NaCl	/	7.2
W3	• • •	II	5 mM Tris-HCl	30 mM NaCl	/	7.2
W2		I	25 mM Tris-HCl	150 mM NaCl	/	7.2
W5	• •	I	25 mM Tris-HCl	1M NaCl	/	7.2
W6	• •	I	5 mM Tris-HCl	1 M NaCl	0.5 mM EDTA	7.5
W4	/			No wash		

4.3.2 dCas9-based enrichment

The dCas9-based enrichment was implemented by designing three new gRNAs (6, 7 and 8, **Table 5**) in close proximity to the *DMPK* microsatellite (**Figure 21**). All gRNAs were first used simultaneously to pull-down the target by forming a RNP complex with biotinylated dCas9. Based on qPCR analysis, average on-target enrichment was 3-fold (**Figure 22D**) and recovered DNA was 10 ng (**Figure 22C**). We reasoned that RNPs in too close proximity may have interfered with each other for target biding. Hence, we tested RNP 7 alone and RNPs 6 and 8 together. Interestingly, RNPs 6 and 8 together provided higher target enrichment (12-fold) as compared to RNP 7 alone (6-fold, **Figure 22D**). A slight increase in recovered target DNA was also reported (15 ng and 16 ng, respectively, **Figure 22C**), but still too low to attempt ONT library preparation. Even though the enrichment performances still need to be improved, these preliminary data showed that RNPs in close proximity (< 300 bp apart) may interfere with each other in target binding. However, a minimum of 2 RNPs seems to be required as starting point for protocol optimization.

4.3.3 DNA-probes based enrichment

For the last approach (DNA-probe-based enrichment), a set of dsDNA-probes was designed to capture a 1.9 kpb region spanning the *DMPK* microsatellite (**Figure 21**). The method yielded the highest fold-enrichment among the three hybridization approaches tested of 537-fold (**Figure 22F**) but the lowest amount of recovered target DNA (1.4 ng, **Figure 22E**). Due to the high costs of the target capture kit, no further test could be performed using DNA-probes.

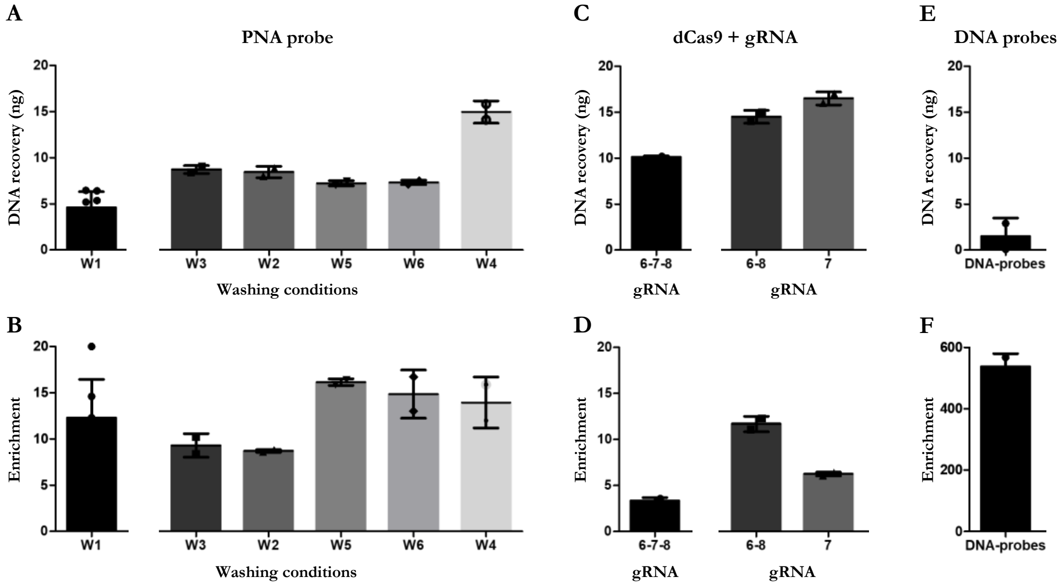


Figure 22. Total recovered DNA and Fold enrichment of hybridization-based capture approaches applied to the *DMPK* microsatellite. (A) Total DNA recovery and **(B)** fold enrichment obtained using the PNA-based approach. Each bar represents the different wash buffers used in the standard condition (W1, N=7) and during the optimization tests (W2-6, N=2) according to Table 3. Wash buffer stringency decreases from left to right. **(C)** Total DNA recovery and **(D)** fold enrichment obtained using the dCas9-based approach. Each bar represents the RNP combinations used in the standard condition (RNPs 6, 7, 8, N=2) and during the optimization tests (RNPs 6, 8 and RNP 7, N=2). **(E)** Total DNA recovery and **(F)** fold enrichment obtained using the ssDNA-based approach (N=2). Due to the high costs of the target capture kit, no further optimization tests could be performed using ssDNA-probes.

Overall, all probe-based enrichment methods produced good-to-optimal target enrichment (12 – 537-fold) at ~1 kbp from the target as determined via qPCR. Enrichment efficiency was also confirmed at a longer range, using primers at 5 kbp from the *DMPK* microsatellite (**Figure 23**). PNA-probes showed comparable enrichment, whereas a slight decrease was observed for dCas9-probes (11- *vs* 7-fold). DNA-probes also showed a lower enrichment at 5 kbp, but still consistently high (182-fold). However, for all methods the recovered target DNA was too low (namely far below 400ng DNA) to attempt any long-read library preparation and sequencing in all cases, even after methods' optimization. For both the PNA- and dCas9- based capture approaches qPCR analysis also showed that most of the target molecules were left on the post-capture supernatant (85% and 55%, respectively), indicating low efficiency of the hybridization step. Post-capture supernatant from the DNA-probe-based workflow was slightly lower and contained 30% of the target molecules. In light of these results, all methods would require further optimization to increase capture efficiency and the amount of recovered DNA, which is a crucial requirement to perform subsequent long-read sequencing.

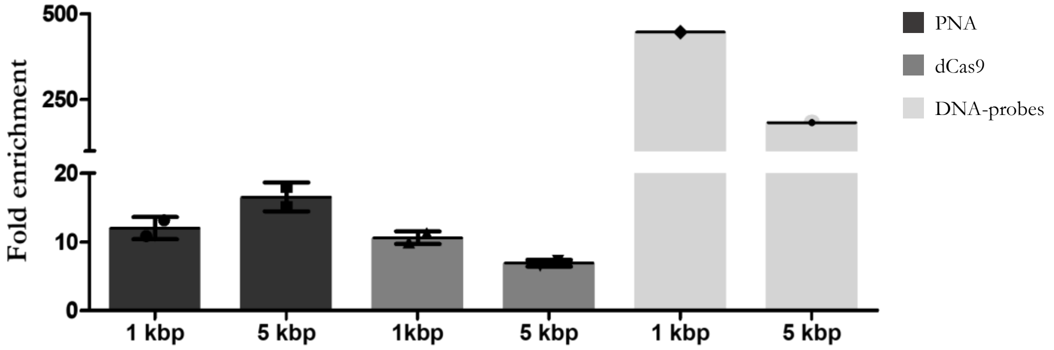


Figure 23. Fold-enrichment of the *DMPK* microsatellite using hybridization-based approaches. Enrichment was evaluated on the optimized protocols for PNA and dCas9 (PNA-W4, dRNP6-8) and on the standard one for DNA-probes. Quantitative PCR was performed using primers at 1 kbp and 5 kbp from the *DMPK* microsatellite, respectively.

4.4 COMPARISON OF LONG-DNA CAPTURE APPROACHES FROM A TECHNICAL POINT OF VIEW

In this section, we compare from a technical point of view the performances of the three long-DNA-fragment enrichment approaches tested (**Table 13**). Xdrop allowed to use the lowest input of gDNA, up to 100 – 1,000 times less as compared to Cas9-mediated capture and hybridization-based approaches. The gDNA extraction method did not significantly influence the enrichment performances of the Xdrop and Cas9-mediated workflows, with the exception of highly viscous DNA (U-HMW). In turn, Cas9-mediated capture was the most sensitive to gDNA quality.

Cas9-mediated capture enabled to achieve the highest fold-enrichment of 662 - 2,467-fold, while the Xdrop indirect capture approach provided just slightly lower target enrichment of 357-fold on average. DNA-probes yielded an enrichment of 537-fold, ~40 times higher as compared to PNA- and dCas9- mediated captures (14 – 12-fold). Xdrop enabled the highest recovery of target DNA (0.8 - 1.3 μg), whereas the same parameter could not be evaluated for Cas9-mediated capture as the excised target was directly sequenced. DNA-probes showed the lowest target recovery (1.2 ng) while PNA- and dCas9 mediated capture allowed higher recovery (up to 15 ng). ONT sequencing could be performed only on DNA enriched using Xdrop- and Cas9-

mediated workflow, generating fragments of ~4.5 kbp and up to 50 kbp, respectively. On the contrary, the length of the region enriched by Xdrop was 100 kbp, twice as long as the maximal size enriched by Cas9-mediated capture (~50 kbp). Enrichment breadth of hybridization-based approaches was successfully validated up to 5 kbp from the *DMPK* microsatellite.

The Xdrop workflow required the use of special instruments, namely a droplet generator with specific cartridges and a flow cytometer, whereas the other approaches only required basic laboratory equipment (thermal cycler and a thermo-block). As a result, the Xdrop workflow was characterized by the highest costs, up to € 1,300/sample when coupled to ONT sequencing. Cas9-mediated capture can only be coupled to ONT sequencing and cost € 1,000 – 1,200 / sample. Hybridization-based approaches provided cheaper assays, with DNA-probes being the cheapest ranging from € 830 to € 50 / sample when coupled to ONT or Illumina sequencing, respectively.

Table 13. Comparison of long-DNA capture approaches from a technical point of view.
Std: standard, HMW: High molecular weight, na: not available

	Xdrop	**Cas9-mediated**	**PNA-probes**	**dCas9-probes**	**DNA-probes**
Enrichment Principle	dPCR, Sorting	Selective adapter ligation	Hybridization	Hybridization	Hybridization
Input gDNA (ng)	10	1,000 - 10,000	5,000	5,000	5,000
Input DNA type	Std to HMW	Std to HMW	Std to HMW	Std to HMW	Std to HMW
Sensitivity to DNA quality	Low	High	na	na	na
Fold enrichment	357	662 - 2,467	14	12	537
DNA recovery (ng)	800 - 1,300	na	15	15	1.2
Length of sequenced DNA fragments (kbp)	4.5	up to 50	na	na	na
Enrichment breadth (kbp)	100	50	Tested up to 5kbp		
Special instruments required	Flow cytometer, Xdrop droplet generator	/	/	/	/
Expected costs (€, Illumina)	600	na	75	80	50
Expected costs (€, ONT)	1,300	1,000 – 1,200	1,000	950	830
Analysis of medium-long repeats	Successful	Successful	Not tested due to low DNA recovery		
Analysis of long repeats	Not possible	Successful			
Analysis of ultra-long repeats	Not possible	Successful			

4.5 COMPARISON OF LONG-DNA CAPTURE APPROACHES

The three long-DNA capture approaches benchmarked in this study showed consistently diverse features and performances both from a technical point of view and for microsatellite expansion analysis.

Cas9-mediated capture enabled to achieve the highest fold-enrichment among the methods tested. Of note, the enrichment yielded by DNA-probes was ~40 times higher as compared to PNA- and dCas9- mediated captures. ONT sequencing of Cas9-enriched samples generated longer fragments as compared to those generated following the Xdrop workflow. On the contrary, the enrichment breadth provided by Xdrop was twice as long as that provided by Cas9-mediated capture.

The analysis of Xdrop-enriched samples by ONT sequencing allowed the accurate assessment of *FMR1* repeat length, along with identification of interruptions and mosaicism. Advantages of the Xdrop workflow were **(a)** the possibility to exploit Illumina sequencing to assess SNVs and Indel in the region surrounding the FMR1 repeat, **(b)** the minimal starting sample requirement, suitable even for small biopsies/pre-natal testing and **(c)** the simplicity of the assay design, comprising just a standard primer set. In turn, the main limitation of the Xdrop method was related to the dMDA step, responsible of shortening the recovered target DNA after amplification, thus not allowing the assessment of contiguous large repeat expansions (>10 kbp). Finally, another drawback is represented by the need of specific instrumentation, making Xdrop the most expensive approach among those tested.

For the analysis of larger repeats, the PCR-free, Cas9-mediated capture coupled to ONT long-read sequencing was more suitable and indeed allowed to assess larger expansions in *CNBP* (up to 46.6 kbp). Importantly, Cas9-mediated capture enabled the accurate identification of repeat length, structure/motif and level of somatic mosaicism. Also, the method did not require any additional instrumentation other than regular molecular biology equipment. On the other hand, the Cas9-method was more sensitive to gDNA quality and indeed it did not allow the successful analysis of DM1 patients from suboptimal samples. Finally, it required much higher DNA input, namely 3 - 5ug.

Hybridization-based approaches were tested only on the *DMPK* microsatellite, that could not be successfully characterized with the previous two approaches. Although all probe-based enrichment methods produced good-to-optimal target enrichment at ~5 kbp, the recovered target DNA was too low to attempt any long-read library preparation and sequencing in all cases, even after methods' optimization. Further optimization is therefore required to increase capture efficiency and the amount of recovered DNA, which is a crucial requirement to perform subsequent long-read sequencing.

5. DISCUSSION

Long-read sequencing has enhanced our capability to characterize genomic regions "dark" to short-reads, harboring large structural variations, repetitive elements, abnormal (>60%) GC content or highly homologous genes (e.g. pseudogenes)[2,3]. The analysis of such genomic features and defects is of utmost importance as they underlie numerous monogenic disorders and complex diseases. However, long-read sequencing approaches still suffer from high costs and to a certain extent also lower accuracy than short-reads. Hence, the combination of long-read sequencing with long-DNA capture approaches can reduce the overall costs while maximizing data production on a selected region of interest. This also enables error-compensation while ensuring sufficient coverage for a more accurate characterization of the target.

Despite their use in combination with long-read sequencing is still poorly explored, several long-DNA-fragment enrichment approaches are available, each based on a very different capture approach. They span from simple hybridization-based approaches to most sophisticated methods involving flow-sorting, restriction enzymes or the CRISPR-Cas9 system. In this thesis, we provided a comparison of different long-DNA capture approaches, namely indirect sequence capture (Samplix's Xdrop), Cas9-mediated targeted sequencing, and a set of three hybridization-capture methods (PNA, dCas9 and dsDNA-probes), taking as case-study three disease causative loci (*FMR1*, *DPMK* and *CNBP*) characterized by microsatellite expansions.

From a technical point of view, on-target enrichment achieved with Xdrop indirect capture and ONT sequencing was very consistent on the *FMR1* locus (357-fold), thus allowing to generate sufficient target coverage to deeply analyze the microsatellite characterizing this gene. Cas9-mediated targeted sequencing enabled to achieve even higher fold-enrichment of 662 to 2,467-fold, in line with previous reports[46,64,68,157]. Such levels of enrichment were instead far to be achieved with probes-based methods, even after protocol optimization. Interestingly, target enrichment was not equal for all target loci. For example, Cas9-mediated enrichment on the *DMPK* microsatellite was ~4 times less as compared to the *CNBP* one. In this case, enrichment efficiency on the *DMPK* locus could have been improved by using >1 gRNA on each flanking sequence in order to increase cut redundancy.

The genome-wide noise of Xdrop analysis (˜0.16x) was lower as compared to that obtained with the Cas9 system coupled to ONT (˜0.32x) in our hands. This indicated that, despite the slightly higher background produced by Cas9-targeted sequencing, the approach allowed indeed to generate higher on-target coverage (446x *vs* 57x). Cas9-targeted sequencing was also attempted using a multiplexing protocol which however generated more background reads than singleplex experiments (0.11x *vs* 0.57x) and consistently lower performances (~10-fold lower enrichment, with 70% unclassified reads). This could be ascribed to the fact that the protocol requires a supplementary step for native barcode ligation. Such procedure possibly determines DNA fragmentation , generation of phosphorylated free-ends that are subsequently ligated by ONT adapters and sequenced, thus producing higher background.

The possibility to capture entire genes and their surroundings can be beneficial for disease diagnosis. Provided good quality HMW DNA, the Xdrop workflow allowed to enrich a region of up to 100 kbp including the whole *FMR1* gene, in agreement with previous reports[36,37]. Such enrichment breadth could be potentially exploited to call SNVs and Indels in the entire gene by coupling the method to Illumina short-read sequencing, the gold standard approach at this aim. This could be useful, for example, for *FMR1*, when the analysis of repeat expansion is inconclusive and the exclusion of other mutations (e.g. SNVs, indels) within the gene is desirable, either to complete genetic testing or to prevent disease transmission[158].

The enrichment breadth provided by the Xdrop workflow could not be achieved with Cas9-mediated capture, because the latter was strictly dependent on the pre-determined target size specified by the gRNA design. Nonetheless, here we demonstrated the enrichment of repeat expansion up to ~50 kbp in length using a single gRNA pair. To our knowledge, the latter represents the longest repeat expansion analyzed so far at the single-nucleotide resolution[45–48] and one of the longest DNA fragment captured using the Cas9-mediated enrichment with no specific adjustment[64,68]. To further expand the enrichment breadth, multiple gRNAs spanning across a longer target region could be design, that allow excising and sequencing multiple overlapping fragments altogether (tiling approach). The possibility to capture significantly long targets by standard workflow (single gRNA pair) or with the "tiling" approach (multiple gRNAs) has important implications for the re-assembly of genomic regions on which short-read data map poorly due to the presence, for instance, of large deletions or repetitive elements.

The performances of the Xdrop and Cas9-mediated capture approaches were tested across a wide set of DNA extraction approaches, spanning from standard to methods yielding U-HMW DNA. Importantly, we found that the choice of the input DNA extraction method did not seem to significantly influence the enrichment performances of the Xdrop and Cas9-mediated workflows, with the exception of highly viscous DNA (U-HMW) which partially interfered with droplet generation or Cas9-mediated cleavage. The possibility to use extraction kits routinely used in diagnostic procedures as well as frozen blood - as our starting samples - could facilitate the broad application of the technologies in the clinic. An exception was the Qiagen gTIP kit that did not properly work in combination with Xdrop. The latter may reflect the carryover of contaminants that interfere with DNA encapsulation/staining and could not be removed using bead-based cleanup methods. In addition, we found that Cas9-mediated capture was more sensitive method to input gDNA quality, as shown by the failure of the Cas9-*DMPK* assay on just partially degraded/impure gDNA from DM1 patients. This was not totally unexpected because this approach is directly coupled (without any intermediate step) to ONT sequencing, whose yield is known to be strongly affected by DNA quality and the presence of impurities[159,160].

One of the advantages of long-DNA capture methods is represented by the possibility to perform PCR-free assays, thus escaping potential biases related to this type of amplification, such as in regions with extreme CG content and repetitions. As a drawback, input gDNA requirements are usually very high (1 – 20 μg) while target recovery can be very low (< 0.01 ng[21]). Indeed, the Cas9-mediated enrichment workflow required the use of 1 – 10 μg input gDNA while up to 5 μg were used for hybridization-based approaches. Some approaches can overcome this limitation starting from lower input and by performing a linear amplification (WGA) on enriched DNA, significantly increasing target recovery up to μg range. This is the case of Xdrop' s indirect capture [36–38], where a specific type of WGA (dMDA) is used following flow-sorting. In our hands, dMDA allowed indeed to recover ~1.1 μg of target DNA on average, providing sufficient substrate both for Illumina and ONT sequencing.

Importantly, the addition of a WGA step downstream the enrichment step also allowed to scale input DNA down to ng range. As demonstrated in this thesis, Xdrop required indeed from 100 – 1,000 times less gDNA as compared to Cas9-mediated capture and hybridization-based approaches (10 ng Vs. 1,000 – 10,000 ng), opening-up to the application of this approach to limited samples such as those derived from pre-natal/pre-implant testing[49], clinical biopsies or even single cells.

Despite the obvious advantage to generate a larger amount of enriched DNA, the multiple displacement mechanism taking place in WGA is known to originate DNA hyperbranches, which in turn can assume many alternative secondary structures. This is happening because the DNA strands extended on an initial template can be displaced becoming available to prime on a second template creating amplification chimeras[148,149,161]. Based on the downstream sequencing approach, DNA must be debranched using either acoustic sonication/ restriction enzymes (short-reads) or the T7 Endonuclease 1 (long-reads). These approaches successfully linearize the DNA but inevitably produce shorter fragments. Despite the this phenomenon is reduced by droplet-based Phi29 amplification[149,162,163] (dMDA) implemented in the Xdrop workflow, sequencing of dMDA-amplified DNA still produces chimeric reads (59.3%), which must be removed during bioinformatic analysis. The removal of chimeric reads can have a detrimental effect on data analysis, as shown in our experiments where the primary alignment length was reduced from 4,098 bp to 1,506 bp, thus limiting the ability to accurately assess repeat lengths expanding over primary alignment sizes. Due to the short length of both enriched DNA and ONT primary alignments, we could indeed not apply the Xdrop workflow to DM1 and DM2 samples, were expansions have been reported to reach up to 19.5 kpb[81] and 44 kbp[78], namely far beyond the length of dMDA amplified DNA. Target length of dMDA-amplified samples may be possibly increased by treating enriched DNA with short-read eliminator (Circulomics), which has been show to deplete short DNA fragments <25 kbp[157,164–166]. On the other hand, decreasing dMDA incubation time may result in longer fragments, reducing however the amount of recovered DNA below 1 μg, and thus limiting subsequent analysis. Performing multiple replicates can overcome the DNA input issue, even though the costs would be significantly higher.

The Xdrop workflow provides an option to assess enrichment by qPCR before proceeding with sequencing, which can represent an advantage to assess the actual experiment outcome before sequencing and thus save costs in case of failure. However, this should be considered solely as a qualitative test to ensure successful results (when > 100x), because there was no full correlation between the enrichment level determined by sequencing and qPCR in our experiments. On the other hand, the Cas9-mediated enrichment workflow does not involve any pull-down step, meaning that the only way to assess gRNA cut efficiency and enrichment would be through sequencing. To reduce risk of failure experiments, we developed a qPCR-based assay to monitor the gRNA cut-efficiency prior to sequencing or to evaluate the performances of a gRNA set. Interestingly, the most efficient gRNAs (>70% cut efficiency) were confirmed via ONT sequencing, demonstrating the utility of our assay for gRNA pre-screening. Quantitative PCR was also used to assess the enrichment efficiency of hybridization-based approaches. DNA-probes provided very good on-target enrichment of 537-fold but yielded the lowest DNA recovery (1.2 ng) that did not allow the subsequent sequencing step, for which a consistently higher amount of DNA (> 400 ng) is required. PNA- and dCas9-based approaches provided a lower enrichment of 14 – 12-fold, accounting for a higher recovery of target DNA as well as higher background. Indeed, target recovery from dCas9- and PNA-mediated capture was consistently higher as compared to previous reports (15 ng *vs* < 0.01 ng[21]), but still too low to attempt ONT sequencing. Interestingly, qPCR analysis showed that most of the target molecules were left on the post-capture supernatant (85% and 55%, respectively), indicating a bias in the hybridization step. On the contrary, post-capture supernatant from the DNA-probe-based workflow only contained 30% of the target molecules. For this reason, further efforts should be focused on the improvement of the hybridization conditions (e.g. buffer composition, temperature, hybridization time), as well as the evaluation of alternative hybridization-based approaches. In this regard, a method involving magnetic beads coated with ssDNA probes has been recently launched on the market (MagIC beads from ElementZero Biolabs, Berlin, Germany). Even though the method has not been yet assessed in literature, according to manufacturers it shows good potential for the PCR-free capture of long-fragments up to 30 kbp with an enrichment of 100,000-fold[167]. In order to bypass the low DNA recovery produced by the hybridization-based approaches, isothermal amplification of the enriched DNA could be performed as demonstrated for the Xdrop workflow. This however would have the drawback of obtaining consistently shorter DNA fragments, thus limiting subsequent analysis, as described above.

In terms of investment required, Xdrop was the most expensive method as it required not only a droplet generator distributed by Samplix but also a flow-sorter including trained personnel for operating it. Samplix is currently developing an all-in-one instrument with the purpose of automatizing the whole workflow, from droplet generation to flow-sorting. This has indeed the potential not only to reduce initial capital costs, but also to further improve the performances by standardizing the workflow. In contrast the Cas9- and hybridization-based methods did not require any dedicated instruments but just standard laboratory equipment. For consumable costs, Cas9- and Xdrop-enrichment coupled to ONT sequencing were very similar and both approaches still resulted very expensive (> € 1,000 / sample). In order to achieve a broader application of these methods, such as in the clinical setting, costs must be reduced. From this point of view, hybridization-based capture approaches could represent a cost-effective alternative to both Xdrop and Cas9-mediated capture workflows, with DNA-probes being the cheapest and ~2 times less expensive when coupled to ONT sequencing. Alternatively, costs could be reduced up to 16 times less if Illumina sequencing is exploited, instead of ONT, an option possible with Xdrop and probes but not with Cas9. While, in order to maintain the benefits of long-read sequencing, ONT cost optimization could be possibly achieved by employing Flongles, that could be acquired at one tenth of the costs of regular ONT flowcells. Even though Flongles are characterized by a lower sequencing amount (1 Gb vs >10Gb of standard flowcell), the latter would be still more than enough for most clinical applications.

Xdrop- and Cas9-mediated workflows enabled successful ONT sequencing of patients harboring known repeat expansions, highlighting their potential for the clinical setting. In particular, Xdrop + ONT allowed the classification of full range *FMR1* alleles (normal, pre-mutation and full mutation), with accurate size estimates comparable to previous results. Furthermore, we could also detect with high-confidence the presence of AGG interruptions, which have been shown to increase repat stability and reduce the risk of expansion in the full mutation range[88,89]. The precise determination of interruption patterns in female (pre-mutation) carriers is therefore critical because it influences their reproductive planning. In addition to repeat interruptions, we also detected a consistent level of mosaicism affecting the size of tandem repeats in pre-mutated and fully mutated alleles. Assessing the variability in CGG repeats within and between tissues is another important aspect of FXS diagnosis because this can influence the clinical phenotype of affected individuals[168].

The choice to use Cas9-mediated capture combined to ONT sequencing was underlined by the need to characterize *CNBP* expansions which are among the longest reported to date (up to 44 kbp). As such, the method allowed to fully characterize *CNBP* normal and expanded alleles at single nucleotide resolution, showing very high concordance with traditional reference approaches in terms of both size and structure. A single incongruence was observed for one expanded allele in patient DM2.B where ONT-sequencing underestimated the size obtained with Southern blot (about 20 *vs* 40 kbp). A possible explanation is the occurrence of DNA damages in the sample analyzed due to the long storage of biobank samples, as for example single-strand nicks. While southern-blotting indeed migrates double-stranded DNA, thus compensating for such issues, ONT-sequencing analyzes single-strand molecules that can be eventually interrupted by nicks (https://community.nanoporetech.com/posts/can-ffpe-repair-be-avoided). Single stranded nicks may be sealed by pre-treating gDNA using a cocktail of enzymes typically used for Formalin-Fixed, Paraffin-Embedded (FFPE) samples. In the expanded alleles of most DM2 patients analyzed (7 out of 9), alongside the expected CCTG repetition, the single-nucleotide resolution has revealed a previously unknown "atypical" repeat with TCTG motif, located at the distal 3'-end of the CCTG array. When present, the motif showed both intra- and inter-donor length variability (40 – 8,000 bp). Possible reasons explaining why such motif has been never reported in DM2 could be either the difficulties in sequencing through the full *CNBP* expanded alleles or the fact that TP-PCR amplification is driven by a primer containing "pure" CCTG repetitions that may not recognize the atypical TCTG motif. Considering that the latter was present in a very variable fraction of expanded alleles (11% to 86%), always in the presence of the typical CCTG repetitions, such technical bias may have thus favored only the amplification of “pure” CCTG repetitions. Our collaborators from the University of Tor Vergata successfully confirmed the atypical TCTG motif via an orthogonal method based on TP-PCR using specific-TCTG primer plus Sanger sequencing. Considering that the “atypical” TCTG motif was discovered in a small set of patients, most of them belonging to the same family, further studies would be required to assess its biological significance. The Cas9-targeted sequencing approach allowed also to estimate the level of somatic mosaicism of *CNBP* mutated alleles, either “pure” or “interrupted”, with intra-patient expansion length variability reaching up 65% on average. Since mosaicism plays an important role in the development of disease symptoms, the determination of the relative percentage of expanded alleles in the lower and upper mutation range could have a prognostic value and significantly improve prognosis and genetic counselling in DM2. Another advantage of the approach utilized is the possibility to perform a PCR-free analysis, that potentially allows the direct assessment of the DNA methylation pattern as already done for other repeat-linked diseases[45,169]. This can constitute an added-on information useful to evaluate the impact of expansions in the functionality of *CNBP* gene.

6. CONCLUSION

In this thesis, we provided a first benchmark of long-DNA capture approaches, identifying and evaluating the strengths and weaknesses of each workflow. Our results could offer a valuable starting point for the widespread application of these technologies, not only in the research setting but also in the clinics. In particular, for the characterization of a wider range of complex genomic regions implicated in different pathogenic conditions and currently poorly/only partially explored by traditional approaches. Accordingly, two of such methods allowed the simultaneous analysis of repeat expansion length, microsatellite structure/motif and level of somatic mosaicism, otherwise not feasible with traditional methods (used either alone or in combination). Comprehensively, these results demonstrated the potential of long-DNA capture approaches to be applied in translational research and in clinical settings, with ultimately strong benefits for the diagnostic workflow and genetic counselling.

7. REFERENCES

1. Ebbert MTW, Jensen TD, Jansen-West K, et al. Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biol.* Published online 2019. doi:10.1186/s13059-019-1707-2
2. Salzberg SL, Yorke JA. Beware of mis-assembled genomes. *Bioinformatics.* Published online 2005. doi:10.1093/bioinformatics/bti769
3. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nat Rev Genet.* Published online 2012. doi:10.1038/nrg3117
4. Porubsky D, Garg S, Sanders AD, et al. Dense and accurate whole-chromosome haplotyping of individual genomes. *Nat Commun.* Published online 2017. doi:10.1038/s41467-017-01389-4
5. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. doi:10.1038/s41576-020-0236-x
6. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* 2020;21(1):1-16. doi:10.1186/s13059-020-1935-5
7. PacBio. Sequencing 101: Understanding Accuracy in DNA Sequencing. Accessed July 2, 2022. https://www.pacb.com/blog/understanding-accuracy-in-dna-sequencing/
8. Hon T, Mars K, Young G, et al. Highly accurate long-read HiFi sequencing data for five complex genomes. doi:10.1038/s41597-020-00743-4
9. Payne A, Holmes N, Rakyan V, Loose M. Bulkvis: A graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics.* 2019;35(13):2193-2198. doi:10.1093/bioinformatics/bty841
10. Jain M, Koren S, Miga KH, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol.* 2018;36(4):338-345. doi:10.1038/nbt.4060
11. Oxford nanopore technology. Oxford Nanopore announces technology updates at Nanopore Community Meeting. https://nanoporetech.com/about-us/news/oxford-nanopore-announces-technology-updates-nanopore-community-meeting
12. Sereika M, Kirkegaard RH, Karst SM, et al. Oxford Nanopore R10.4 long-

read sequencing enables near-perfect bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *bioRxiv*. Published online 2021:2021.10.27.466057. https://www.biorxiv.org/content/10.1101/2021.10.27.466057v2%0Ahttps://www.biorxiv.org/content/10.1101/2021.10.27.466057v2.abstract

13. Sekhar C, Chilamakuri R, Lorenz S, et al. Performance comparison of four exome capture systems for deep sequencing. *BMC Genomics*. 2014;15(449):1-13.

14. Clark MJ, Chen R, Lam HYK, et al. Performance comparison of exome DNA sequencing technologies. 2011;29. doi:10.1038/nbt.1975

15. Shigemizu D, Momozawa Y, Abe T, et al. Performance comparison of four commercial human whole-exome capture platforms OPEN. Published online 2015. doi:10.1038/srep12742

16. Altmüller J, Motameny S, Becker C, et al. A systematic comparison of two new releases of exome sequencing products: The aim of use determines the choice of product. *Biol Chem*. 2016;397(8):791-801. doi:10.1515/hsz-2015-0300

17. Iadarola B, Xumerle L, Lavezzari D, et al. Shedding light on dark genes: enhanced targeted resequencing by optimizing the combination of enrichment technology and DnA fragment length. Published online 2020. doi:10.1038/s41598-020-66331-z

18. Iadarola B, Lavezzari D, Modi A, et al. Whole-exome sequencing of the mummified remains of Cangrande della Scala (1291-1329 CE) indicates the first known case of late-onset Pompe disease. *Sci Reports* |. 123AD;11:21070. doi:10.1038/s41598-021-00559-1

19. Slesarev A, Viswanathan L, Tang Y, et al. CRISPR/Cas9 targeted CAPTURE of mammalian genomic regions for characterization by NGS. *Sci Rep*. Published online 2019. doi:10.1038/s41598-019-39667-4

20. Chandler DP, Stults JR, Anderson KK, Cebula S, Schuck BL, Brockman FJ. Affinity capture and recovery of DNA at femtomolar concentrations with peptide nucleic acid probes. *Anal Biochem*. Published online 2000. doi:10.1006/abio.2000.4637

21. Murphy NM, Pouton CW, Irving HR. Human leukocyte antigen haplotype

phasing by allele-specific enrichment with peptide nucleic acid probes. *Mol Genet Genomic Med.* Published online 2014. doi:10.1002/mgg3.65

22. Nielsen PE, Egholm M, Berg RH, Buchardt O. Sequence-selective recognition of DNA by strand displacement with a thymine-substituted polyamide. *Science (80-).* 1991;254(5037):1497-1500. doi:10.1126/science.1962210

23. Jensen KK, Ørum H, Nielsen PE, Nordén B. *Kinetics for Hybridization of Peptide Nucleic Acids (PNA) with DNA and RNA Studied with the BIAcore Technique †.* Vol 36.; 1997. https://pubs.acs.org/sharingguidelines

24. Qi LS, Larson MH, Gilbert LA, et al. Repurposing CRISPR as an RNA-γuided platform for sequence-specific control of gene expression. *Cell.* 2013;152(5):1173-1183. doi:10.1016/j.cell.2013.02.022

25. Ma H, Tu LC, Naseri A, et al. Multiplexed labeling of genomic loci with dCas9 and engineered sgRNAs using CRISPRainbow. *Nat Biotechnol.* 2016;34(5):528-530. doi:10.1038/nbt.3526

26. Anton T, Karg E, Bultmann S. Applications of the CRISPR/Cas system beyond gene editing. doi:10.1093/biomethods/bpy002

27. Bethune K, Mariac C, Couderc M, et al. Long-fragment targeted capture for long-read sequencing of plastomes. *Appl Plant Sci.* 2019;7(5):1-13. doi:10.1002/aps3.1243

28. Dapprich J, Ferriola D, Mackiewicz K, et al. The next generation of target capture technologies - large DNA fragment enrichment and sequencing determines regional genomic variation of high complexity. *BMC Genomics.* Published online 2016. doi:10.1186/s12864-016-2836-6

29. Lippow SM, Aha PM, Parker MH, Blake WJ, Baynes BM, Lipovš ek D. Creation of a type IIS restriction endonuclease with a long recognition sequence. *Nucleic Acids Res.* 2009;37(9):3061-3073. doi:10.1093/nar/gkp182

30. Bath AJ, Milsom SE, Gormley NA, Halford SE. Many type IIs restriction endonucleases interact with two recognition sites before cleaving DNA. *J Biol Chem.* 2002;277(6):4024-4033. doi:10.1074/jbc.M108441200

31. Pham TT, Yin J, Eid JS, et al. Single-locus enrichment without amplification for sequencing and direct detection of epigenetic modifications. *Mol Genet Genomics.* Published online 2016. doi:10.1007/s00438-016-1167-2

32. Hommelsheim CM, Frantzeskakis L, Huang M, Bekir &. PCR amplification of

repetitive DNA: a limitation to genome editing technologies and many other applications. Published online 2014. doi:10.1038/srep05052

33. Kanagawa T. Bias and artifacts in multitemplate polymerase chain reactions (PCR). *J Biosci Bioeng*. 2003;96(4):317-323. doi:10.1016/s1389-1723(03)90130-7

34. Day DJ, Speiser PW, Schulze E, et al. *Identification of Non-Amplifying CYP21 Genes When Using PCR-Based Diagnosis of 21-Hydroxylase Deficiency in Congenital Adrenal Hyperplasia (CAH) Affected Pedigrees*. Vol 5.; 1996. https://academic.oup.com/hmg/article/5/12/2039/658375

35. Ogino S, Wilson RB. Quantification of PCR Bias caused by a single nucleotide polymorphism in SMN gene dosage analysis. *J Mol Diagnostics*. 2002;4(4):185-190. doi:10.1016/S1525-1578(10)60702-7

36. Madsen EB, Höijer I, Kvist T, Ameur A, Mikkelsen MJ. Xdrop: Targeted sequencing of long DNA molecules from low input samples using droplet sorting. *Hum Mutat*. Published online 2020. doi:10.1002/humu.24063

37. Blondal T, Gamba C, Møller Jagd L, et al. Verification of CRISPR editing and finding transgenic inserts by Xdrop indirect sequence capture followed by short- and long-read sequencing. *Methods*. 2021;191(May 2020):68-77. doi:10.1016/j.ymeth.2021.02.003

38. Grosso V, Marcolungo L, Maestri S, et al. Characterization of FMR1 Repeat Expansion and Intragenic Variants by Indirect Sequence Capture. 2021;12(September). doi:10.3389/fgene.2021.743230

39. Dejesus-Hernandez M, Aleff RA, Jackson JL, et al. Long-read targeted sequencing uncovers clinicopathological associations for C9orf72-linked diseases. Published online 2021. doi:10.1093/brain/awab006

40. Ebbert MTW, Farrugia S, Sens J, et al. Long-read sequencing across the C9orf72 'GGGGCC' repeat expansion: implications for clinical use and genetic discovery efforts in human disease. *bioRxiv*. Published online 2018:1-17. doi:10.1101/176651

41. Hafford-Tear NJ, Tsai YC, Sadan AN, et al. CRISPR/Cas9-targeted enrichment and long-read sequencing of the Fuchs endothelial corneal dystrophy–associated TCF4 triplet repeat. *Genet Med*. 2019;21(9):2092-2102. doi:10.1038/s41436-019-0453-x

42. Höijer I, Tsai YC, Clark TA, et al. Detailed analysis of HTT repeat elements in

human blood using targeted amplification-free long-read sequencing. *Hum Mutat*. 2018;39(9):1262-1272. doi:10.1002/humu.23580

43. Tsai YC, Greenberg D, Powell J, et al. Amplification-free, CRISPR-Cas9 targeted enrichment and SMRT sequencing of repeat-expansion disease causative genomic regions. *bioRxiv*. Published online 2017:1-26. doi:10.1101/203919

44. Wieben ED, Aleff RA, Basu S, et al. Amplification-free long-read sequencing of TCF4 expanded trinucleotide repeats in Fuchs Endothelial Corneal Dystrophy. *PLoS One*. 2019;14(7):1-14. doi:10.1371/journal.pone.0219446

45. Giesselmann P, Brändl B, Raimondeau E, et al. Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. *Nat Biotechnol*. 2019;37(12):1478-1481. doi:10.1038/s41587-019-0293-x

46. Mizuguchi T, Toyota T, Miyatake S, et al. Complete sequencing of expanded SAMD12 repeats by long-read sequencing and Cas9-mediated enrichment . *Brain*. Published online 2021:1-14. doi:10.1093/brain/awab021

47. Sone J, Mitsuhashi S, Fujita A, et al. Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. *Nat Genet*. 2019;51(8):1215-1221. doi:10.1038/s41588-019-0459-y

48. Wallace AD, Sasani TA, Swanier J, et al. CaBagE: A Cas9-based Background Elimination strategy for targeted, long-read DNA sequencing. *PLoS One*. 2021;16(4):e0241253. doi:10.1371/journal.pone.0241253

49. Mosca-Boidron A-L, Faivre L, Aho S, Marle N, Truntzer C. An Improved Method to Extract DNA from 1 ml of Uncultured Amniotic Fluid from Patients at Less than 16 Weeks' Gestation. *PLoS One*. 2013;8(4):59956. doi:10.1371/journal.pone.0059956

50. Kuderna LFK, Lizano E, Julià E, et al. Selective single molecule sequencing and assembly of a human Y chromosome of African origin. *Nat Commun*. Published online 2019. doi:10.1038/s41467-018-07885-5

51. K Kuderna LF, Solís-Moruno M, Batlle-Masó L, et al. Flow Sorting Enrichment and Nanopore Sequencing of Chromosome 1 From a Chinese Individual. doi:10.3389/fgene.2019.01315

52. Brouns SJJ, Jore MM, Lundgren M, et al. Small CRISPR RNAs Guide

Antiviral Defense in Prokaryotes. *Proc Natl Acad Sci USA*. 1990;1(6):8448. doi:10.1126/science.1103388

53. Makarova KS, Grishin N V, Shabalina SA, Wolf YI, Koonin E V. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. doi:10.1186/1745-6150-1-7

54. Mojica FJM, Díez-Villaseñor C, García-Martínez J, Almendros C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology*. Published online 2009. doi:10.1099/mic.0.023960-0

55. Makarova KS, Haft DH, Barrangou R, et al. *Evolution and Classification of the CRISPR–Cas Systems.*; 2011. doi:10.1038/nrmicro2577

56. Makarova KS, Aravind L, Wolf YI, Koonin E V. *Unification of Cas Protein Families and a Simple Scenario for the Origin and Evolution of CRISPR-Cas Systems.* Vol 6.; 2011. doi:10.1186/1745-6150-6-38

57. Deltcheva E, Chylinski K, Sharma CM, et al. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III Europe PMC Funders Group. *Nature*. 2011;471(7340):602-607. doi:10.1038/nature09886

58. Jinek M, Jiang F, Taylor DW, et al. Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science (80-)*. 2014;343(6176). doi:10.1126/science.1247997

59. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. *A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity*. https://www.science.org

60. Gasiunas G, Barrangou R, Horvath P, Siksnys V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. doi:10.1073/pnas.1208507109

61. Sternberg SH, Redding S, Jinek M, et al. HHHHS Public Access Author manuscript Nature. Author manuscript; available in PMC 2014 September 06. Published in final edited form as: Nature. 2014 March 6; 507(7490): 62–67. doi:10.1038/nature13011. DNA interrogation by the CRISPR RNA-guided endonucleas. *Nature*. 2014;507(7490):62-67. doi:10.1038/nature13011.DNA

62. Scherer S. *A Short Guide to the Human Genome*. Cold Spring Harbor Laboratory

Press; 2008.

63. Hryhorowicz M, Lipiński D, Zeyland J, Słomski R. CRISPR/Cas9 Immune System as a Tool for Genome Engineering. *Arch Immunol Ther Exp (Warsz)*. 2017;65(3):233-240. doi:10.1007/s00005-016-0427-5

64. Gilpatrick T, Lee I, Graham JE, et al. Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat Biotechnol*. Published online 2020. doi:10.1038/s41587-020-0407-5

65. Hsu PD, Lander ES, Zhang F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell*. Published online 2014. doi:10.1016/j.cell.2014.05.010

66. Jiang W, Zhao X, Gabrieli T, Lou C, Ebenstein Y, Zhu TF. Cas9-Assisted Targeting of CHromosome segments CATCH enables one-step targeted cloning of large gene clusters. *Nat Commun*. Published online 2015. doi:10.1038/ncomms9101

67. Gabrieli T, Sharim H, Fridman D, Arbib N, Michaeli Y, Ebenstein Y. Selective nanopore sequencing of human BRCA1 by Cas9-assisted targeting of chromosome segments (CATCH). *Nucleic Acids Res*. 2018;46(14):87. doi:10.1093/nar/gky411

68. Iyer S V, Kramer M, Goodwin S, Mccombie WR. ACME: an Affinity-based Cas9 Mediated Enrichment method for targeted nanopore sequencing. *bioRxiv*. Published online 2022. doi:10.1101/2022.02.03.478550

69. López-Girona E, Davy MW, Albert NW, et al. CRISPR-Cas9 enrichment and long read sequencing for fine mapping in plants. *Plant Methods*. 2020;16(1):1-13. doi:10.1186/s13007-020-00661-x

70. Gymrek M. A genomic view of short tandem repeats. *Curr Opin Genet Dev*. 2017;44:9-16. doi:10.1016/j.gde.2017.01.012

71. Willems T, Gymrek M, Highnam G, Mittelman D, Erlich Y. The landscape of human STR variation. *Genome Res*. 2014;24(11):1894-1904. doi:10.1101/gr.177774.114

72. Cumming SA, Hamilton MJ, Robb Y, et al. De novo repeat interruptions are associated with reduced somatic instability and mild or absent clinical features in myotonic dystrophy type 1. *Eur J Hum Genet*. 2018;26:1635-1647. doi:10.1038/s41431-018-0156-9

73. Pearson CE, Eichler EE, Lorenzetti D, et al. *Interruptions in the Triplet Repeats of SCA1 and FRAXA Reduce the Propensity and Complexity of Slipped Strand DNA (S-DNA) Formation †.*; 1998. https://pubs.acs.org/sharingguidelines

74. Kraus-Perrotta C, Lagalwar S. Expansion, mosaicism and interruption: mechanisms of the CAG repeat mutation in spinocerebellar ataxia type 1. doi:10.1186/s40673-016-0058-y

75. Botta A, Visconti VV, Fontana L, et al. A 14-Year Italian Experience in DM2 Genetic Testing: Frequency and Distribution of Normal and Premutated CNBP Alleles. *Front Genet.* 2021;12. doi:10.3389/fgene.2021.668094

76. Santoro M, Masciullo M, Silvestri G, Novelli G, Botta A. Myotonic dystrophy type 1: role of CCG, CTC and CGG interruptions within DMPK alleles in the pathogenesis and molecular diagnosis. *Clin Genet.* 2017;92(4):355-364. doi:10.1111/cge.12954

77. Charles P, Camuzat A, Benammar N, et al. Are interrupted SCA2 CAG repeat expansions responsible for parkinsonism? *Neurology.* 2007;69(21):1970-1975. doi:10.1212/01.wnl.0000269323.21969.db

78. Depienne C, Mandel JL. 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? *Am J Hum Genet.* 2021;108(5):764-785. doi:10.1016/j.ajhg.2021.03.011

79. Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. *Nat Rev Genet.* 2018;19(5):286-298. doi:10.1038/nrg.2017.115

80. Brown V, Warren ST. Trinucleotide Repeats: Dynamic DNA and Human Disease. *Brenner's Encycl Genet Second Ed.* Published online January 1, 2001:181-185. doi:10.1016/B978-0-12-374984-0.01580-1

81. Lanni S, Pearson CE. Molecular genetics of congenital myotonic dystrophy. *Neurobiol Dis.* 2019;132(March):104533. doi:10.1016/j.nbd.2019.104533

82. Krieger K, Eichler EE, Holden JJA, et al. Length of uninterrupted CGG repeats determines instability in the FMR1 gene. Published online 1994. Accessed February 10, 2022. http://www.nature.com/naturegenetics

83. Nolin SL, Glicksman A, Ersalesi N, et al. Fragile X full mutation expansions are inhibited by one or more AGG interruptions in premutation carriers. Published online 2014. doi:10.1038/gim.2014.106

84. Yrigollen CM, Martorell L, Durbin-Johnson B, et al. AGG interruptions and

maternal age affect FMR1 CGG repeat allele stability during transmission. *J Neurodev Disord*. Published online 2014. doi:10.1186/1866-1955-6-24

85. Saldarriaga W, Carrera G, Tassone F, et al. *Fragile X Syndrome Síndrome de X Frágil*. Vol 45.; 2014.

86. Bardoni B, Schenck A, Mandel JL. The Fragile X mental retardation protein. *Brain Res Bull*. 2001;56(3-4):375-382. doi:10.1016/S0361-9230(01)00647-5

87. Hagerman RJ, Berry-Kravis E, Hazlett HC, et al. Fragile X syndrome. *Nahum Sonenb*. 2017;1. doi:10.1038/nrdp.2017.65

88. Nolin SL, Brown WT, Glicksman A, et al. Expansion of the fragile X CGG repeat in females with premutation or intermediate alleles. *Am J Hum Genet*. 2003;72(2):454-464. doi:10.1086/367713

89. Yrigollen CM, Durbin-Johnson B, Gane LM, et al. AGG interruptions within the maternal FMR1 gene reduce the risk of offspring with fragile X syndrome. Published online 2012. doi:10.1038/gim.2012.34

90. Matsuyama Z, Izumi Y, Kameyama M, Kawakami H, Nakamura S. The eVect of CAT trinucleotide interruptions on the age at onset of spinocerebellar ataxia type 1 (SCA1). doi:10.1136/jmg.36.7.546

91. Charles P, Camuzat A, Benammar N, et al. GGA·TCC-interrupted Triplets in Long GAA·TTC Repeats Inhibit the Formation of Triplex and Sticky DNA Structures, Alleviate Transcription Inhibition, and Reduce Genetic Instabilities. *J Biol Chem*. 2007;276(21):1970-1975. doi:10.1074/jbc.M101852200

92. Tabolacci E, Pietrobono R, Maneri G, et al. Reversion to Normal of FMR1 Expanded Alleles: A Rare Event in Two Independent Fragile X Syndrome Families. doi:10.3390/genes11030248

93. van Blitterswijk M, DeJesus-Hernandez M, Niemantsverdriet E, et al. Association between repeat sizes and clinical and pathological characteristics in carriers of C9ORF72 repeat expansions (Xpansize-72): A cross-sectional cohort study. *Lancet Neurol*. 2013;12(10):978-988. doi:10.1016/S1474-4422(13)70210-2

94. Quartier A, Poquet H, Gilbert-Dussardier B, et al. Intragenic FMR1 disease-causing variants: a significant mutational mechanism leading to Fragile-X syndrome. *Nat Publ Gr*. 2017;25:423-431. doi:10.1038/ejhg.2016.204

95. Loomis EW, Eid JS, Peluso P, et al. Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile X gene. *Genome Res.* 2013;23(1):121-128. doi:10.1101/GR.141705.112

96. Thornton CA. Myotonic dystrophy. *Neurol Clin.* 2014;32(3):705-719. doi:10.1016/j.ncl.2014.04.011

97. Bird TD. Myotonic Dystrophy Type 1 Summary Genetic counseling Suggestive Findings. *GeneReviews®.* Published online 2019:1-27. https://www.ncbi.nlm.nih.gov/books/NBK1165/pdf/Bookshelf_NBK1165.pdf

98. Brook JD, McCurrach ME, Harley HG, et al. Molecular basis of myotonic dystrophy: Expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member. *Cell.* 1992;68(4):799-808. doi:10.1016/0092-8674(92)90154-5

99. Fu YH, Pizzuti A, Fenwick RG, et al. An unstable triplet repeat in a gene related to myotonic muscular dystrophy. *Science (80-).* 1992;255(5049):1256-1258. doi:10.1126/science.1546326

100. Mahadevan M, Tsilfidis C, Sabourin L, et al. Myotonic dystrophy mutation: An unstable CTG repeat in the 3' untranslated region of the gene. *Science (80-).* 1992;255(5049):1253-1255. doi:10.1126/science.1546325

101. Murányi A, Zhang R, Liu F, et al. Myotonic dystrophy protein kinase phosphorylates the myosin phosphatase targeting subunit and inhibits myosin phosphatase activity. *FEBS Lett.* 2001;493(2-3):80-84. doi:10.1016/S0014-5793(01)02283-9

102. Kaliman P, Llagostera E. Myotonic dystrophy protein kinase (DMPK) and its role in the pathogenesis of myotonic dystrophy 1. *Cell Signal.* 2008;20(11):1935-1941. doi:10.1016/j.cellsig.2008.05.005

103. Cho DH, Tapscott SJ. Myotonic dystrophy: Emerging mechanisms for DM1 and DM2. *Biochim Biophys Acta - Mol Basis Dis.* 2007;1772(2):195-204. doi:10.1016/j.bbadis.2006.05.013

104. Tsilfidis C, MacKenzie AE, Mettler G, Barceló J, Korneluk RG. Correlation between CTG trinucleotide repeat length and frequency of severe congenital myotonic dystrophy. *Nat Genet.* 1992;1(3):192-195. doi:10.1038/ng0692-192

105. López Castel A, Cleary JD, Pearson CE. *Repeat Instability as the Basis for Human*

Diseases and as a Potential Target for Therapy.; 2010. doi:10.1038/nrm2854

106. Musova Z, Mazanec R, Krepelova A, et al. Highly unstable sequence interruptions of the CTG repeat in the myotonic dystrophy gene. *Am J Med Genet A*. 2009;149A(7):1365-1374. doi:10.1002/ajmg.a.32987

107. Braida C, Stefanatos RKA, Adam B, et al. Variant CCG and GGC repeats within the CTG expansion dramatically modify mutational dynamics and likely contribute toward unusual symptoms in some myotonic dystrophy type 1 patients. *Hum Mol Genet*. 2010;19(8):1399-1412. doi:10.1093/hmg/ddq015

108. Santoro M, Masciullo M, Pietrobono R, et al. Molecular, clinical, and muscle studies in myotonic dystrophy type 1 (DM1) associated with novel variant CCG expansions. *J Neurol*. 2013;260(5):1245-1257. doi:10.1007/s00415-012-6779-9

109. Botta A, Rossi G, Marcaurelio M, et al. Identification and characterization of 5' CCG interruptions in complex DMPK expanded alleles. *Eur J Hum Genet*. 2017;25(2):257-261. doi:10.1038/ejhg.2016.148

110. Pešović J, Perić S, Brkušanin M, Brajušković G, Rakoč Ević -Stojanović V, Savić-Pavić Ević D. Repeat interruptions modify age at onset in myotonic dystrophy type 1 by stabilizing DMPK expansions in somatic cells. *Front Genet*. Published online 2018. doi:10.3389/fgene.2018.00601

111. Lian M, Zhao M, Lee CG, Chong SS. Single-Tube Dodecaplex PCR Panel of Polymorphic Microsatellite Markers Closely Linked to the DMPK CTG Repeat for Preimplantation Genetic Diagnosis of Myotonic Dystrophy Type 1. Published online 2017. doi:10.1373/clinchem.2017.271528

112. Mcvicker GP, Breuss MW, Chong SS, Lian M, Lee CG. Robust Preimplantation Genetic Testing Strategy for Myotonic Dystrophy Type 1 by Bidirectional Triplet-Primed Polymerase Chain Reaction Combined With Multi-microsatellite Haplotyping Following Whole-Genome Amplification. *Artic 589 Genet*. 2019;10:589. doi:10.3389/fgene.2019.00589

113. Meola G, Cardani R. Myotonic Dystrophy Type 2: An Update on Clinical Aspects, Genetic and Pathomolecular Mechanism. *J Neuromuscul Dis*. 2015;2:59-71. doi:10.3233/JND-150088

114. Montagnese F, Mondello S, Wenninger S, Kress W, Benedikt Schoser ·. Assessing the influence of age and gender on the phenotype of myotonic

dystrophy type 2. *J Neurol.* 1234;264:2472-2480. doi:10.1007/s00415-017-8653-2

115. Liquori CL, Ricker K, Moseley ML, et al. Myotonic dystrophy type 2 caused by a CCTG expansion in intron I of ZNF9. *Science (80-).* 2001;293(5531):864-867. doi:10.1126/science.1062125

116. Benhalevy D, Gupta SK, Danan CH, et al. The Human CCHC-type Zinc Finger Nucleic Acid-Binding Protein Binds G-Rich Elements in Target mRNA Coding Sequences and Promotes Translation. *Cell Rep.* 2017;18(12):2979-2990. doi:10.1016/J.CELREP.2017.02.080

117. Radvanszky J, Surovy M, Polak E, Kadasi L. Uninterrupted CCTG tracts in the myotonic dystrophy type 2 associated locus. *Neuromuscul Disord.* 2013;23(7):591-598. doi:10.1016/j.nmd.2013.02.013

118. Mahyera AS, Schneider T, Halliger-Keller B, et al. Distribution and structure of DM2 Repeat tract alleles in the German population. *Front Neurol.* Published online 2018. doi:10.3389/fneur.2018.00463

119. Guo P, Lam SL. Unusual structures of CCTG repeats and their participation in repeat expansion. *Biomol Concepts.* 2016;7(5-6):331-340. doi:10.1515/bmc-2016-0024

120. Kamsteeg E-J, Kress W, Catalli C, et al. Best practice guidelines and recommendations on the molecular diagnosis of myotonic dystrophy types 1 and 2. Published online 2012. doi:10.1038/ejhg.2012.108

121. Udd B, Meola G, Krahe R, et al. Report of the 115th ENMC workshop: DM2/PROMM and other myotonic dystrophies: 3rd Workshop, 14-16 February 2003, Naarden, The Netherlands. In: *Neuromuscular Disorders.* Vol 13. Elsevier Ltd; 2003:589-596. doi:10.1016/S0960-8966(03)00092-0

122. Day JW, Ricker K, Jacobsen JF, et al. Myotonic dystrophy type 2: Molecular, diagnostic and clinical spectrum. *Neurology.* 2003;60(4):657-664. doi:10.1212/01.WNL.0000054481.84978.F9

123. Bachinski LL, Udd B, Meola G, et al. Confirmation of the Type 2 Myotonic Dystrophy (CCTG)n Expansion Mutation in Patients with Proximal Myotonic Myopathy/Proximal Myotonic Dystrophy of Different European Origins: A Single Shared Haplotype Indicates an Ancestral Founder Effect. *Am J Hum Genet.* 2003;73(4):835-848. doi:10.1086/378566

124. Spector E, Behlmann A, Kronquist K, Rose NC, Lyon E, Reddi H V. Laboratory testing for fragile X, 2021 revision: a technical standard of the American College of Medical Genetics and Genomics (ACMG). *Genet Med.* 2021;23(5):799-812. doi:10.1038/s41436-021-01115-y

125. Paulson H. Repeat expansion diseases. Published online 2018. doi:10.1016/B978-0-444-63233-3.00009-9

126. Lockhart PJ. Advancing the diagnosis of repeat expansion disorders. Published online 2022. doi:10.1016/S1474-4422(22)00028-X

127. Botta A, Bonifazi E, Vallo L, et al. Italian guidelines for molecular analysis in myotonic dystrophies. *Acta Myol myopathies cardiomyopathies Off J Mediterr Soc Myol.* 2006;25(1):23-33.

128. Gu H, Kim MJ, Yang D, et al. Accuracy and Performance Evaluation of Triplet Repeat Primed PCR as an Alternative to Conventional Diagnostic Methods for Fragile X Syndrome. *Ann Lab Med.* 2021;41(4):394-400. doi:10.3343/alm.2021.41.4.394

129. Rajan-Babu IS, Chong SS. Triplet-repeat primed PCR and capillary electrophoresis for characterizing the fragile X mental retardation 1 CGG repeat hyperexpansions. *Methods Mol Biol.* 2019;1972:199-210. doi:10.1007/978-1-4939-9213-3_14

130. Radvansky J, Ficek A, Minarik G, Palffy R, Kadasi L. Effect of unexpected sequence interruptions to conventional PCR and repeat primed PCR in myotonic dystrophy type 1 testing. *Diagn Mol Pathol.* 2011;20(1):48-51. doi:10.1097/PDM.0b013e3181efe290

131. German society of human genetics. Indication Criteria for Genetic Testing. Published online 2008.

132. Monaghan KG, Lyon E, Spector EB. ACMG standards and guidelines for fragile X testing: A revision to the disease-specific supplements to the standards and guidelines for Clinical Genetics Laboratories of the American College of Medical Genetics and Genomics. *Genet Med.* 2013;15(7):575-586. doi:10.1038/gim.2013.61

133. Warner JP, Barron LH, Goudie D, et al. A general method for the detection of large CAG repeat expansions by fluorescent PCR. *JMed Genet.* 1996;33(2):1022-1026. doi:10.1136/jmg.33.12.1022

134. Adler K, Moore JK, Filippov G, Wu S, Carmichael J, Schermer M. A novel assay for evaluating fragile X locus repeats. *J Mol Diagnostics*. 2011;13(6):614-620. doi:10.1016/j.jmoldx.2011.06.002

135. Bastepe M, Xin W. Huntington Disease: Molecular Diagnostics Approach. *Curr Protoc Hum Genet*. 2015;87(1):9.26.1-9.26.23. doi:10.1002/0471142905.hg0926s87

136. Hayward BE, Zhou Y, Kumari D, Usdin K. A Set of Assays for the Comprehensive Analysis of FMR1 Alleles in the Fragile X–Related Disorders. *J Mol Diagnostics*. 2016;18(5):762-774. doi:10.1016/j.jmoldx.2016.06.001

137. Saluto A, Brussino A, Tassone F, et al. An enhanced polymerase chain reaction assay to detect pre- and full mutation alleles of the fragile X mental retardation 1 gene. *J Mol Diagnostics*. 2005;7(5):605-612. doi:10.1016/S1525-1578(10)60594-6

138. Filipovic-Sadic S, Sah S, Chen L, et al. A Novel FMR1 PCR Method for the Routine Detection of Low Abundance Expanded Alleles and Full Mutations in Fragile X Syndrome. Published online 2009. doi:10.1373/clinchem.2009.136101

139. Ardui S, Race V, Ravel T de, et al. Detecting AGG interruptions in females with a FMR1 premutation by long-read single-molecule sequencing: A 1 year clinical experience. *Front Genet*. Published online 2018. doi:10.3389/fgene.2018.00150

140. Mcfarland KN, Liu J, Landrian I, et al. SMRT Sequencing of Long Tandem Nucleotide Repeats in SCA10 Reveals Unique Insight of Repeat Expansion Structure. Published online 2015. doi:10.1371/journal.pone.0135906

141. Wiley HS, Chen R, Ameur A, Hoischen A, Mantere T, Kersten S. Long-Read Sequencing Emerging in Medical Genetics. Published online 2019. doi:10.3389/fgene.2019.00426

142. Mangin A, de Pontual L, Tsai YC, et al. Robust detection of somatic mosaicism and repeat interruptions by long-read targeted sequencing in myotonic dystrophy type 1. *Int J Mol Sci*. 2021;22(5):1-24. doi:10.3390/ijms22052616

143. Ciosi M, Cumming SA, Chatzi A, et al. Approaches to Sequence the HTT CAG Repeat Expansion and Quantify Repeat Length Variation. *J Huntingtons*

Dis. 2021;10(1):53-74. doi:10.3233/JHD-200433

144. Riet J, Ramos LR V, Lewis R V, Marins LF, Mol G. Improving the PCR protocol to amplify a repetitive DNA sequence. *Genet Mol Res*. 2017;16(3):16039796. doi:10.4238/gmr16039796

145. Chakraborty S, Vatta M, Bachinski LL, Krahe R, Dlouhy S, Bai S. Molecular diagnosis of myotonic dystrophy. *Curr Protoc Hum Genet*. 2016;2016(October):9.29.1-9.29.19. doi:10.1002/cphg.22

146. De Coster W, D'hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. doi:10.1093/bioinformatics/bty149

147. Li H. Minimap2: pairwise alignment for nucleotide sequences. doi:10.1093/bioinformatics/bty191

148. Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. Published online 2016. doi:10.1038/nrg.2015.16

149. Zhou X, Xu Y, Zhu L, et al. Comparison of Multiple Displacement Amplification (MDA) and Multiple Annealing and Looping-Based Amplification Cycles (MALBAC) in Limited DNA Sequencing Based on Tube and Droplet. doi:10.3390/mi11070645

150. Benson G. *Tandem Repeats Finder: A Program to Analyze DNA Sequences*. Vol 27.; 1999. https://academic.oup.com/nar/article/27/2/573/1061099

151. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. *Integrative Genomics Viewer*.; 2011. doi:10.1038/nbt0111-24

152. Chen S, Zhou Y, Chen Y, Gu J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018;34(17):i884-i890. doi:10.1093/bioinformatics/bty560

153. Maestri S, Maturo MG, Cosentino E, et al. A Long-Read Sequencing Approach for Direct Haplotype Phasing in Clinical Settings. *Int J Mol Sci Artic*. doi:10.3390/ijms21239177

154. Wilson JA, Pratt VM, Phansalkar A, et al. Consensus characterization of 16 FMR1 reference materials: A consortium study. *J Mol Diagnostics*. 2008;10(1):2-12. doi:10.2353/jmoldx.2008.070105

155. Lim GXY, Yeo M, Koh YY, et al. Validation of a commercially available test that enables the quantification of the numbers of CGG trinucleotide repeat

expansion in FMR1 gene. Published online 2017. doi:10.1371/journal.pone.0173279

156. Lin YC, Boone M, Meuris L, et al. Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. *Nat Commun*. 2014;5. doi:10.1038/NCOMMS5767

157. Kirov I, Polkhovskaya E, Dudnikov M, et al. Searching for a needle in a haystack: Cas9-targeted nanopore sequencing and dna methylation profiling of full-length glutenin genes in a big cereal genome. *Plants*. 2022;11(1). doi:10.3390/plants11010005

158. Sitzmann AF, Robert |, Hagelstrom T, Flora Tassone |, Hagerman RJ, Butler MG. Rare FMR1 gene mutations causing fragile X syndrome: A review. Published online 2017. doi:10.1002/ajmg.a.38504

159. Vaillancourt B, Buell CR. High molecular weight DNA isolation method from diverse plant species for use with Oxford Nanopore sequencing. *bioRxiv*. Published online 2019:783159. https://doi.org/10.1101/783159

160. Brown RB, Wass T, Sudhakar N, Brown R. Efficient NGS ready gDNA from microalga. Published online 2019:1-9.

161. Lasken RS, Stockwell TB. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol*. 2007;7:1-11. doi:10.1186/1472-6750-7-19

162. Gonzalez-Pena V, Natarajan S, Xia Y, et al. Accurate genomic variant detection in single cells with primary template-directed amplification. doi:10.1073/pnas.2024176118/-/DCSupplemental

163. Hård J, Mold JE, Eisfeldt J, et al. Long-read whole genome analysis of human single cells. *bioRxiv*. Published online 2021:2021.04.13.439527. https://doi.org/10.1101/2021.04.13.439527

164. Auxier B, Becker F, Nijland R, Debets AJM, Van Den Heuvel J, Snelders E. Meiosis in the human pathogen Aspergillus fumigatus has the highest known number of crossovers. *bioRxiv*. Published online 2022. doi:10.1101/2022.01.14.476329

165. Dicke SS, Dustin Rubinstein C, Speers JM, et al. A protocol for locating and counting transgenic sequences from laboratory animals using a map-then-capture (MapCap) sequencing workflow: procedure and application of results.

bioRxiv. Published online 2022. doi:10.1101/2022.01.13.476149

166. Fukunaga K, Abe A, Mukainari Y, et al. Recombinant inbred lines and next-generation sequencing enable rapid identification of candidate genes involved in morphological and agronomic traits in foxtail millet. *Sci Rep*. Published online 2022. doi:10.1038/s41598-021-04012-1

167. ElementZero Biolabs. ElementZero. Published 2022. Accessed February 23, 2022. https://elementzero.bio/product/custom-target-dna-seq-magic-beads/

168. Pretto D, Yrigollen CM, Tang H-T, et al. Clinical and molecular implications of mosaicism in FMR1 full mutations. Published online 2014. doi:10.3389/fgene.2014.00318

169. Fukuda H, Yamaguchi D, Nyquist K, et al. Father-to-offspring transmission of extremely long NOTCH2NLC repeat expansions with contractions: genetic and epigenetic profiling with long-read sequencing. *Clin Epigenetics*. 2021;13(1):1-17. doi:10.1186/s13148-021-01192-5