Resolving challenging medically-relevant genes using nanopore sequencing

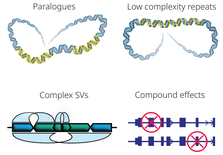
Sergey Nurk[1], Alexey Zabelkin[1], Lynn Ly[2], Heather Jeffery[1], Rebecca Stubbs[3], Bryan Leland[3], Carly Tyer[2], Sean McKenzie[2], Phil James[1], Sissel Juul[3], Philipp Rescheneder[1]

[1]Oxford Nanopore Technologies, plc Oxford, UK
[2]Oxford Nanopore Technologies, Inc. San Francisco, CA
[3]Oxford Nanopore Technologies, Inc. New York City, NY

Abstract

Variant calling has become a routine component of genomic analysis, enabling the identification of single-nucleotide variants, small insertions and deletions, and structural variants across the human genome. However, a subset of challenging medically-relevant genes (CMRGs) remains difficult to analyze. Examples include gene/pseudogene pairs embedded in large segmental duplications, as well as genes overlapping complex repeat structures such or genes overlapping short tandem repeats. In segmental duplications, high sequence identity among paralogs complicates read mapping and paralog-specific variant interpretation, while frequent gene conversions, and gene–pseudogene hybrids further obscure copy-specific analysis when using short reads. At repeat-expansion loci, motif length, sequence composition, and haplotype context often determine pathogenicity, yet are inaccessible to short-read methods. Oxford Nanopore long-read sequencing addresses these limitations by spanning segmental duplications, low-complexity repeats, and providing long-range phasing in a single assay. It also preserves DNA methylation and supports both reference-guided and assembly-based approaches. Using reference cell lines with known pathogenic variants, we demonstrate that Nanopore sequencing provides a robust and flexible platform for interrogating CMRGs, enabling accurate variant calling, haplotype resolution, and improved diagnostic insight in regions previously considered intractable.

1. Introduction



Challenging medically-relevant genes (CMRGs) harbor clinically actionable variation in sequence contexts that are inaccessible to short-read sequencing[1]. CMRG categories include multi-copy genes arising from large, highly identical segmental duplications; repeat-expansion and short tandem repeat loci; regions enriched for complex structural variation or complex repeats; and regions of low sequence complexity. Clinical interpretation frequently requires establishing cis/trans phase for compound-heterozygous variants tens to hundreds of kilobases apart. Oxford Nanopore long reads span repetitive regions, enable phasing, and, where required, support local or whole-genome assembly to resolve complex structural variation. Furthermore, preservation of methylation information renders specialised assays unnecessary at loci where epigenetic state contributes to pathogenicity.

2. Resolving short tandem repeats

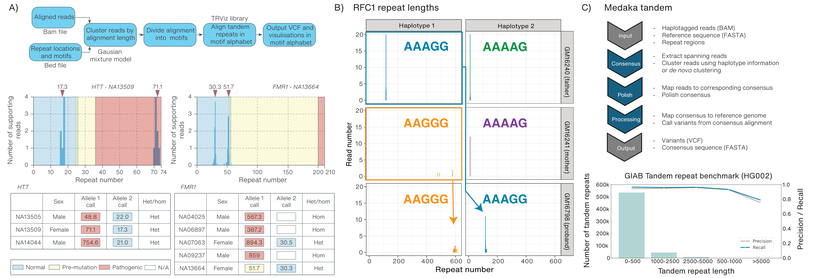
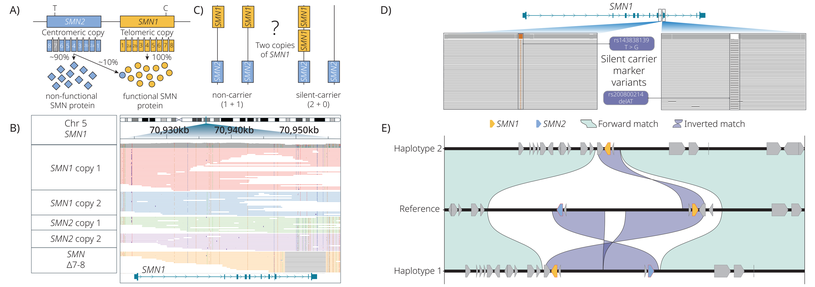


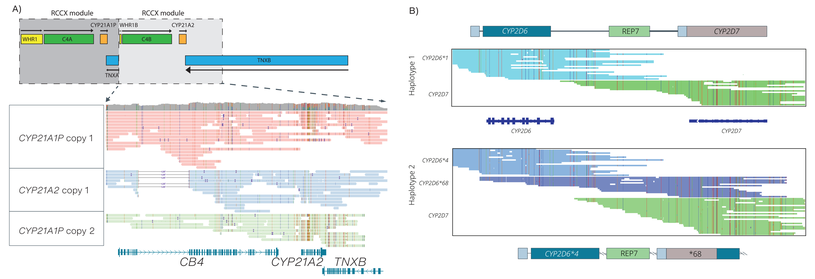
Figure 2A illustrates allele- and haplotype-specific analysis of repeat expansions at clinically relevant short tandem repeats (STRs). Individual reads span the repeat as well as unique flanks, enabling direct sizing of each allele and unambiguous phasing to parental haplotypes. We demonstrated locus-specific repeat length estimation by reporting read-level repeat counts, using *wf-human-variation*, for Huntington's disease (*HTT*) across three samples and for Fragile-X-Syndrome (*FMR1*) across five samples, capturing both normal and expanded alleles as well as inter-read variability. However, pathogenicity at several loci depends not only on length but also on repeat composition[2]. To illustrate nanopore's ability to distinguish between disease-associated and non-pathogenic motifs we sequenced a trio with repeats expansions in *RFC1* (Fig. 2B). Moving beyond locus-by-locus assays, *medaka tandem* is a scalable tool for genome-wide analysis of STR lengths across a large catalog of repeats (Fig. 2C). We applied this method to the *addoto* repeat catalogue[3] and evaluated performance against the GIAB STR benchmark, observing high precision and recall.

3. Full characterisation of the SMA locus



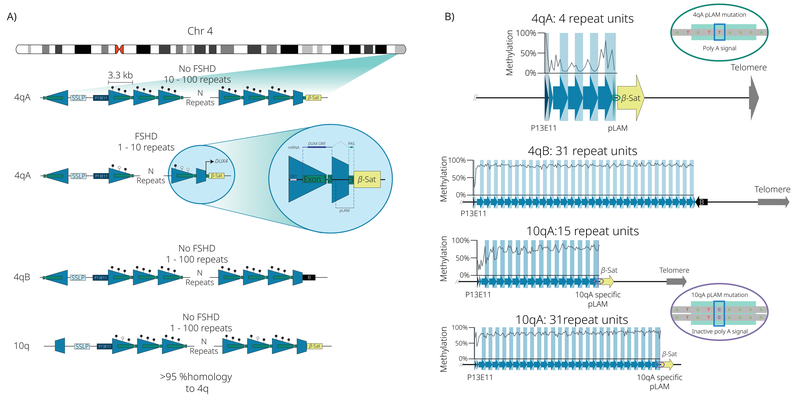
Spinal muscular atrophy results from loss-of-function variants in, or complete loss of, *SMN1*[4] (Fig. 3A). Accurate identification of these variants is challenging because the near-identical paralog *SMN2* confounds read mapping, copy-number estimation, and variant attribution with short reads. Here, we demonstrate that, by spanning paralogous regions and informative paralog-specific bases, Nanopore sequencing enables a reference-based phasing approach assigning reads directly to *SMN1* or *SMN2* haplotypes (Fig. 3B). This approach yields haplotype-resolved *SMN1* and *SMN2* copy numbers and detects pathogenic SNPs, indels, and structural variants in each gene. However, a major remaining challenge is the identification of silent carriers i.e. individuals with a haplotype lacking *SMN1* (Fig. 3C). Traditionally, silent-carrier risk is assessed using marker variants, but these proxies are imperfect[5]. Figure 3D shows that the two most common markers are present in the GM2029 cell line, suggesting silent-carrier status. Only through genome assembly (Fig. 3E) were we able to uncover that, in spite of the marker variants, both haplotypes contain a copy of *SMN1*.

4. Resolving paralogues and multi-copy genes



Resolving paralogues and multi-copy genes requires discriminating near-identical copies, assigning variants and copy number to the correct paralogue and haplotype, and the detection of gene conversions or rearrangements. As summarised in the SMA panel above, long-read, reference-based phasing successfully resolves *SMN1* and *SMN2*. This approach extends to other clinically important CMRGs, including *PMS2/PMS2CL*, *HBA1/HBA2* and *CYP21A2/CYP21A1P*. As a representative case, we resolved a copy-number loss of *CYP21A2* within the RCCX module (Fig. 4A). However, some CMRGs, such as *CYP2D6*, require full haplotype reconstruction even in the presence of complex hybrids. To resolve *CYP2D6*, we used chinook, an assembly-based analysis tool that reconstructs individual gene copies, resolves hybrid junctions, and differentiates *CYP2D6* from *CYP2D7 (Fig. 4B)*. Together, reference-guided phasing and targeted assembly deliver paralogue- and haplotype-specific variant calls, robust copy-number estimates, and clinically interpretable haplotypes across diverse multi-copy genes.

5. Macro-satellites and complex repeat structures



Facioscapulohumeral muscular dystrophy (FSHD) results from the misexpression of the *DUX4* gene encoded within the *D4Z4* array located on chromosome 4[6]. Misexpression of stable mRNA from *DUX4* is caused by aberrant chromatin architecture and subsequent hypo-methylation due to a contraction of the D4Z4 array to under 10 repeat units. There are two dominant alleles in the population, termed 4qA and 4qB, which exist at approximately equal frequencies. Due to the presence of a pLAM site and associated polyA signal (PAS), FSHD is only observed in individuals with a contracted and/or hypo-methylated 4qA allele. Contractions and methylation status of the 4qB allele, or a region on chromosome 10 that shares approximately 98% homology with the 4q region, have no effect on the FSHD phenotype (Fig. 5A). The cell line GM17939 was obtained from an individual with FSHD. Using *hifiasm*[7] with presets optimised for Oxford Nanopore sequencing, the 4q and 10q regions were assembled and used as input into the DZ4Z repeat caller, *D4Z4End2End*[8]. Raw reads were mapped back to the assemblies and methylation status of the whole array was enumerated. The 4qA allele showed a significant contraction to within the pathogenic range and the array was shown to be hypo-methylated. The 4qA specific pLAM and polyA signal required for production of stable mRNA was identified in the assembly. The contig representing a 4qB allele was identified using a B specific sequence located towards the telomeric region of the array, while the 10q specific pLAM site was identified in the assembled haplotypes representing the 10q homologous regions (Fig. 5B).

Conclusion

Oxford Nanopore sequencing enables the comprehensive characterization of challenging medically-relevant genes (CMRGs) that remain inaccessible to short-read sequencing. Across diverse genomic contexts, including segmental duplications, repeat expansions, and epigenetically regulated regions Nanopore sequencing enables phasing of variants, resolving of structural rearrangements, quantification of allele-specific copy number, and detection of clinically relevant methylation states. By combining reference-based phasing, assembly-based reconstruction, and methylation-aware analysis, we demonstrate accurate interpretation for genes such as *SMN1*, *CYP2D6*, *FMR1*, *RFC1*, and *DUX4*. These results highlight the value of long-read sequencing as a scalable and unified platform for variant detection in CMRGs, supporting both research and translational applications.

References

1. Ji Y, Zhao J, Gong J, Sedlazeck FJ, Fan S. Unveiling novel genetic variants in 370 challenging medically relevant genes using the long read sequencing data of 41 samples from 19 global populations. Mol Genet Genomics. 2024;299(1):65
2. Dominik N, Magri S, Currò R, et al. Normal and pathogenic variation of RFC1 repeat expansions: implications for clinical diagnosis. Brain. 2023;146(12):5060-5069. doi:10.1093/brain/awad240
3. English AC, Dolzhenko E, Ziaei Jam H, et al. Analysis and benchmarking of small and large genomic variants across tandem repeats. Nat Biotechnol. 2025;43(3):431-442. doi:10.1038/s41587-024-02225-z
4. Farrar MA, Kiernan MC. The Genetics of Spinal Muscular Atrophy: Progress and Challenges. Neurotherapeutics. 2015;12(2):290-302. doi:10.1007/s13311-014-0314-x
5. Chen X, Sanchis-Juan A, French CE, et al. Spinal muscular atrophy diagnosis and carrier screening from genome sequencing data. Genet Med. 2020;22(5):945-953. doi:10.1038/s41436-020-0754-0
6. van der Maarel SM, Tawil R, Tapscott SJ. Facioscapulohumeral muscular dystrophy and DUX4: breaking the silence. Trends Mol Med. 2011;17(5):252-258. doi:10.1016/j.molmed.2011.01.001
7. Cheng H, Asri M, Lucas J, Koren S, Li H. Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. Nat Methods. 2024;21(6):967-970. doi:10.1038/s41592-024-02269-8
8. https://www.medrxiv.org/content/10.1101/2025.04.24.25326320v1