


Long read sequencing enhances pathogenic and novel variation discovery in patients with rare diseases

Received: 18 April 2024

Accepted: 28 February 2025

Published online: 14 March 2025

 Check for updates

Shruti Sinha^{1,5} , Fatma Rabea^{2,5}, Sathishkumar Ramaswamy^{1,5}, Ikram Chekroun², Maha El Naofal¹, Ruchi Jain¹, Roudha Alfalasi¹, Nour Halabi¹, Sawsan Yaslam¹, Massomeh Sheikh Hassani¹, Shruti Shenbagam¹, Alan Taylor¹, Mohammed Uddin², Mohamed A. Almarri^{2,3}, Stefan Du Plessis², Alawi Alsheikh-Ali² & Ahmad Abou Tayoun^{1,4} 

With ongoing improvements in the detection of complex genomic and epigenomic variations, long-read sequencing (LRS) technologies could serve as a unified platform for clinical genetic testing, particularly in rare disease settings, where nearly half of patients remain undiagnosed using existing technologies. Here, we report a simplified funnel-down filtration strategy aimed at enhancing the identification of small and large deleterious variants as well as abnormal episignature disease profiles from whole-genome LRS data. This approach detected all pathogenic single nucleotide, structural, and methylation variants in a positive control set ($N = 76$) including an independent sample set with known methylation profiles ($N = 57$). When applied to patients who previously had negative short-read testing ($N = 51$), additional diagnoses were uncovered in 10% of cases, including a methylation profile at the spinal muscular atrophy locus utilized for diagnosing this life-threatening, yet treatable, condition. Our study illustrates the utility of LRS in clinical genetic testing and the discovery of novel disease variation.

Around 7000 rare diseases have been identified, collectively imposing significant health and socio-economic burden¹. The majority of these diseases have a genetic origin due to variants ranging from single nucleotide variants (SNVs) or a few nucleotide insertions/deletions (INDELs) to large genomic changes such as copy number variants (CNVs), translocations, inversions, transposable element (TE) insertions, or complex rearrangements. Some are also associated with specific epigenomic profiles². This diverse spectrum of disease-causing changes, often detected by different technologies, has challenged current genetic diagnostic strategies and contributed to long diagnostic odysseys, averaging at 6 years³, delaying timely management or treatment plans for patients with rare diseases.

Although short-read sequencing technologies have brought a remarkable leap in the diagnosis of rare genetic diseases^{4,5}, more than half of the patients remain undiagnosed. This is partly due to the inherent limitations of this technology in detecting complex variants such as structural variants, methylation profiles, repeat expansions, or variants embedded in inaccessible regions of the genome, specifically high homology and GC-rich regions⁶. Recent advances in third-generation sequencing technologies have demonstrated the application of targeted LRS for identifying pathogenic variants in known or novel disease-causing genes^{7–10}. However, the clinical implementation of LRS for detecting genome-wide variation and methylation changes in the context of rare diseases has been limited by challenges

¹Dubai Health Genomic Medicine Center, Dubai Health, Dubai, UAE. ²Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai Health, Dubai, UAE. ³Genome Center, Department of Forensic Science and Criminology, Dubai Police GHQ, Dubai, UAE. ⁴Center for Genomic Discovery, Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai Health, Dubai, UAE. ⁵These authors contributed equally: Shruti Sinha, Fatma Rabea, Sathishkumar Ramaswamy. ✉ e-mail: ajch_ssinha@dubaihealth.ae; Ahmad.Tayoun@dubaihealth.ae

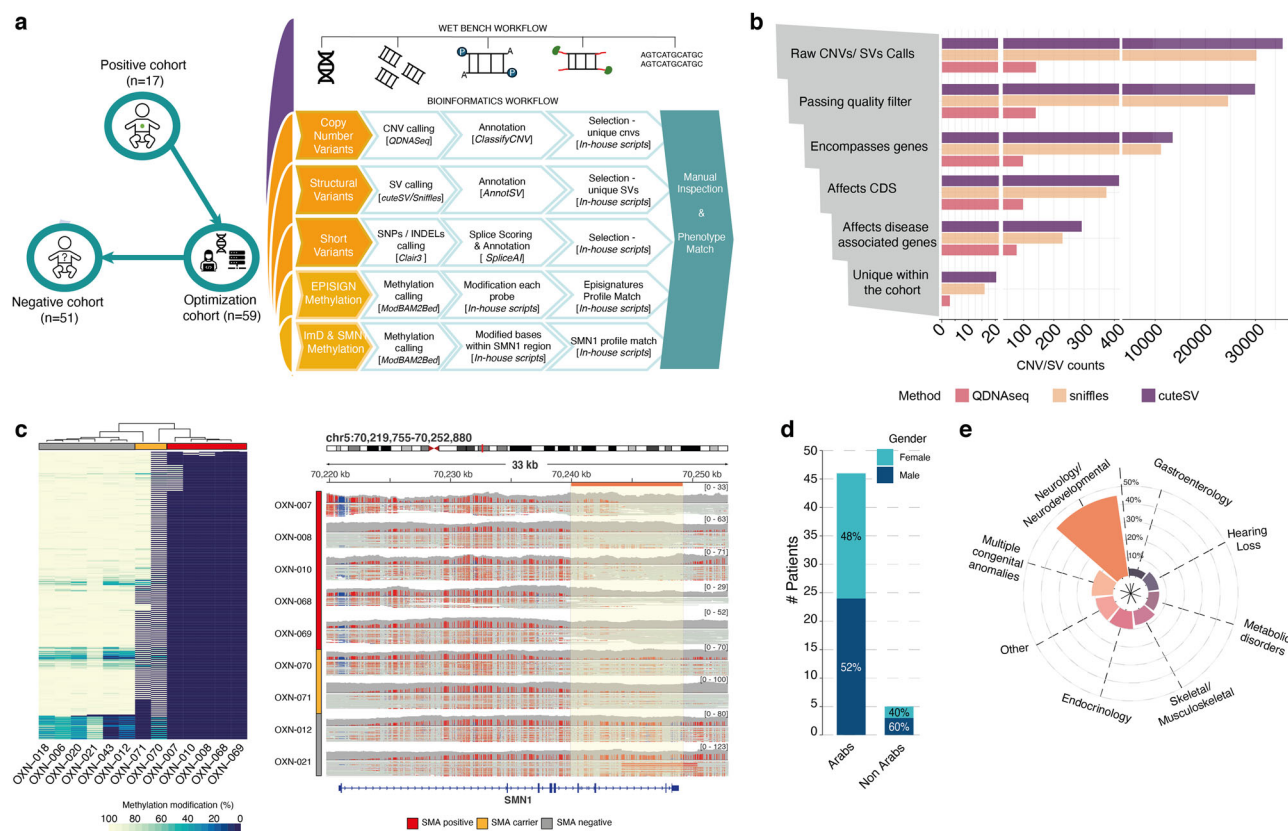


Fig. 1 | Study design, proof of concept in positive cohort and overview of negative cohort. Shown are (a) the study design schema, along with the wet bench and bioinformatics workflows with the tools used. Information about the positive ($N=17$) and negative ($N=51$) samples are in Supplementary Fig. 1a and Supplementary Data 2. Samples used for optimization ($N=59$) include two samples with *SMN1* carrier status and 57 samples set with previously confirmed methylation profiles, as explained in the main text. (b) counts of CNVs and SVs for each method in

each filtering step of the “funnel-down” approach. (c) aggregate methylation profile of SMA –heatmap of base methylation modification (%) within the *SMN1* chr5:70239954-70249165 region (left panel) and IGV methylation view in *SMN1* (right panel) for SMA positive (OXN-007, OXN-008, OXN-010, OXN-068, OXN-069), SMA carrier (OXN-070, OXN-071) and SMA negative samples (OXN-012, OXN-021). (d) the gender and demography and (e) the most prevalent primary clinical symptom in the “Negative Cohort”. Source data are provided as a Source Data file.

associated with the annotation and filtration of a large number of variants and is, therefore, yet to be explored. Here, we optimize a whole genome LRS workflow and a filtration strategy which we apply in a cohort of undiagnosed patients with suspected rare diseases leading to additional diagnoses and the uncovering of a methylation signature utilized for the diagnosis of Spinal Muscular Atrophy (SMA).

Results and discussion

We optimized our analysis workflow on a selected cohort of 17 patients with confirmed genetic diagnoses, encompassing a diverse array of genomic and epigenomic pathogenic variants (Fig. 1a and Supplementary Fig. 1a). The study design incorporated wet bench protocol optimized for long-read Oxford Nanopore sequencing using PromethION system targeting a minimum of 30X coverage with average N50 of 12 kb (Fig. 1a and Supplementary Fig. 1b). Our computational analysis workflow consists of a “genome” and “epigenome” modules (Fig. 1a and Supplementary Methods). The former module consists of detection, annotation, and selection of short variants (SNVs and INDELs) (below) and genome-wide rearrangements, mainly copy number variations (CNVs) and structural variations (SVs). Raw variants were retained if calls were supported by ≥ 5 reads with allele fraction ≥ 0.3 and were affecting the coding region of genes associated with disease as defined in OMIM or GeneCC (Supplementary Methods). This reduced the number of called variants by 58.8% for CNVs and 99.2% for SVs. Further filtering of variants unique to each patient in the cohort reduced CNVs by 97.3% (average $n=3$) and SVs to 99.9% (average $n=46$) (Fig. 1b and Supplementary Fig. 1c), which were then manually

inspected for any clinical correlation. This led to the detection of all associated pathogenic variants in this group (Supplementary Fig. 1d–f and Supplementary Fig. 2a, b).

The epigenomic module “Epimarker” scans epismarkers specific to 36 Mendelian neurodevelopmental disorders (MNDD)², including Angelman syndrome and Spinal Muscular Atrophy (SMA) (see below). This module correctly detected loss of methylation at 15q11.2 in a patient with Angelman syndrome (OXN-18), while SNP analysis revealed substantial loss of heterozygosity (LOH) across chromosome 15, suggesting paternal uniparental disomy (pUPD) as the underlying mechanism of disease in this patient (Supplementary Fig. 1e, f). We also optimized Epimarker using an independent LRS dataset ($N=57$) comprising 9 MNDD in a cohort of 17 patients, with clinically confirmed abnormal methylation profiles, along with 40 controls¹⁰. Epimarker classified all patients with 100% sensitivity, while none of the control samples were assigned to MNDD (100% specificity) (Supplementary Data 1 and Supplementary Methods). We further observed a methylation profile across the *SMN1* gene, whose biallelic loss causes spinal muscular atrophy (SMA), and its homologous pseudogene, *SMN2*, and explored the utilization of this methylation pattern as a diagnostic marker for the disease (Fig. 1c) We observed 0–15% (low), 50–70% (moderate) and 98–100% (high) of bases with methylation modification for SMA patients ($N=5$), carriers ($N=2$) and non-carriers ($N=3$), respectively, in the locus spanning intron 6, exon 7 and intron 8 of the *SMN1* gene (chr5:70239954-70249165) (Fig. 1c and Supplementary Fig. 2a). SMA is a common, life-threatening autosomal recessive neuromuscular disease mostly caused by biallelic deletion of exon 7 in

*SMN1*¹¹. We investigated this finding by deconvoluting reads in *SMN1* and *SMN2* based on 16 paralog-specific variants (PSVs)^{12,13} (Supplementary Methods and Supplementary Fig. 2b). Upon deconvolution, we confirmed the biallelic loss of *SMN1* in SMA patients, where no specific reads were mapped to this gene; this finding was further confirmed by droplet digital PCR (Supplementary Fig. 2c). Notably, SMA carriers had a reduced number of reads in *SMN1* relative to non-carriers (Supplementary Fig. 2b–d). Taken together, we propose a workflow where methylation spanning *SMN1* introns 6 and 8 (including exon 7) can be used as a “tag” for SMA diagnosis and carrier status determination, which can then be confirmed upon LRS read deconvolution using *SMN1* and *SMN2* PSVs at this locus (Supplementary Fig. 2b). Overall, our pipeline was able to correctly identify all the pathogenic variants, including complex rearrangements and aberrant methylation, in the optimization cohort.

We applied this workflow to a set of undiagnosed patients ($N = 51$), who previously had inconclusive testing using short read whole exome sequencing (WES). Among them, 41% had undertaken multiple genetic testing, of which 86% had received chromosomal microarray (CMA) testing (Fig. 1a and Supplementary Data 2). Patients were mostly of Arab descendant (90%), had overall equal gender representation (~44% females), and primarily presented with neurological disorders (45%) (Fig. 1d, e and Supplementary Data 2). Whole genome LRS in this cohort obtained an average of 49X coverage and N50 of 11.7Kb (Supplementary Fig. 1b). Since all the samples were previously tested by WES, we focused on SNVs within exons that were missed by WES and those in the 50 bp exon-intron boundary by assessing their splicing potential (Supplementary Data 3). We detected ~47,000 LRS-specific exonic and ~41,000 splicing SNVs comprising 1.8% of the total detected SNVs for each patient (Supplementary Data 3). We applied our SNV filtration criteria (Supplementary Fig. 3a) resulting in approximately ~2 LRS-specific exonic and ~5 splicing SNVs for manual inspection in accordance with ACMG guidelines¹⁴ (Supplementary Methods and Supplementary Fig. 3b). We then evaluated variants within genes associated with diseases matching patients' phenotypes and identified a single variant in *DNMT1* (NM_001130823: c.891 + 8 C > T) in OXN-044 with a highly predicted splicing impact (SpliceAI score = 0.93). Indeed, transcriptomic sequencing using RNA extracted from peripheral blood in this patient confirmed a splicing defect whereby the c.891 + 8 C > T variant introduced a cryptic donor splice site leading to intronic retention of six nucleotides (Supplementary Fig. 3c and Supplementary Data 4, 5 and 6). However, this change introduced two in-frame amino acids and is therefore unlikely to affect protein function as corroborated by the high allele frequency of the c.891 + 8 C > T variant in the general population and specifically in the Middle East (6.5% allele frequency with 25 homozygotes in gnomAD v4.1.0). Therefore, this variant was classified as clinically benign. No other putative clinically relevant sequence variants were identified.

We next focused on larger genomic rearrangements and detected ~35,000 SVs and ~83 CNVs in each sample (Supplementary Data 3). These were substantially reduced by 98.5% and 99.9%, respectively, after applying our filtering and selection criteria (Fig. 2a, Supplementary Fig. 3a and Supplementary Data 3). Within filtered large CNV events, we identified pathogenic variants in two patients. For patient OXN-033, two deletions from a total of 59 CNVs were prioritized, of which a heterozygous deletion event (1.4 Mb) at 2q11.1-q11.2 was classified as pathogenic post manual inspection, was validated by CMA and found to be de novo upon parental testing (Fig. 2b and Supplementary Data 7). Individuals with 2q11.2 deletions have developmental delay, intellectual disability, dysmorphic features, and variable skeletal anomalies along with obesity¹⁵ which was consistent with this patient's phenotype. The other prioritized heterozygous deletion in this patient was 1.4 Mb in size at 15q13.1-q13.2 as confirmed by CMA. The only known disease-causing gene in this region is *NSMCE3* which has been associated with autosomal recessive immunodeficiency and lung

disease (MIM# 617241). Therefore, this deletion, confirmed to be paternally inherited, was not considered to be diagnostic in this patient. In another patient (OXN-048), with unconfirmed diagnosis of anterior segment dysgenesis and a heterozygous pathogenic variant in the *SLC38A8* gene identified by exome sequencing, we detected a single heterozygous deletion (80 kb) at 16q23.3 (Fig. 2c and Supplementary Data 7), partially encompassing *SLC38A8* (exons 8 – 3'UTR), using LRS. *SLC38A8* is associated with autosomal recessive foveal hypoplasia and/or anterior segment dysgenesis matching the patient's phenotype¹⁶. Taking advantage of the long reads, we phased the two variants and observed that each variant is in a distinct haplotype confirming the compound heterozygous configuration in this individual and the biallelic impairment of *SLC38A8* (Fig. 2c).

We then examined the landscape of structural variants. We identified a homozygous deletion of 3.6 kb detected by both Sniffles and CuteSV, partially including the 3' untranslated region (UTR) of the gene encoding the M-Phase Specific PLK1 Interacting Protein (*MPLKIP*) in patient OXN-027 (Fig. 2d and Supplementary Data 7). This patient showed signs of learning disabilities with distinctive brittle hair, a hallmark of Trichothiodystrophy non-photosensitive 1 associated with non-functional MPLKIP protein. The 3'UTR region is known to regulate mRNA-based processes¹⁷, hence we hypothesized that the homozygous 3'UTR deletion of the *MPLKIP* gene could alter its expression levels. In fact, transcriptomic analysis, using RNA extracted from peripheral blood in this patient, showed that this gene is significantly overexpressed (Fig. 2d) in this patient suggesting that its dysregulation might underlie the observed phenotype, especially in light of the MPLKIP protein role as a cell cycle and mitosis regulator where its function might be dependent on its expression levels. However, further investigation is required to confirm this mechanism and to understand the functional impact of the 3'UTR deletion in this gene.

We next used “Epimarker” to compare the methylation patterns of the 51 undiagnosed patients with the epismutation profiles associated with 34 MNDD². One patient (OXN-062) was classified, based on its methylation profile, as having Hunter McAlpine syndrome (HMA) with a duplication at 5q35.2-q35.3 containing the *NSD1* gene detected by LRS CNV analysis and validated by chromosomal microarrays (Fig. 2e and Supplementary Data 7). HMA is characterized by craniosynostosis, intellectual deficit, short stature, and facial dysmorphism matching the clinical indication of the patient. While deletions of *NSD1* and hypomethylation at this locus are associated with Sotos syndrome, HMA has been associated with micro-duplication involving *NSD1* and a hyper-methylation profile² confirming the diagnosis for this patient. We then examined the SMA methylation tag described above across all undiagnosed patients. Interestingly, we observed a methylation profile consistent with biallelic loss of *SMN1* in the patient (OXN-060). Reads deconvolution analysis as well as droplet digital PCR confirmed SMA diagnosis in this patient who also had 4 copies of the *SMN2* gene (Fig. 2f).

The protocols for analyzing LRS are still in nascent stages, and no global standard methods have been established. In this study, we aimed to establish a comprehensive diagnostic workflow for LRS to investigate the genomic and epigenomic landscape in rare disorders focusing on variants disrupting disease-causing genes or loci. We propose a filtering strategy which substantially reduces the number of variants detected by whole genome LRS while capturing a wide spectrum of genomic and epigenomic pathogenic variation, leading to 10% (5 out of 51) additional diagnoses in patients with rare diseases who had inconclusive testing using traditional methods. We acknowledge that our approach might have reduced sensitivity for several reasons, including the possibility of filtering out large SV events in non-coding regions, which might still be causative through several

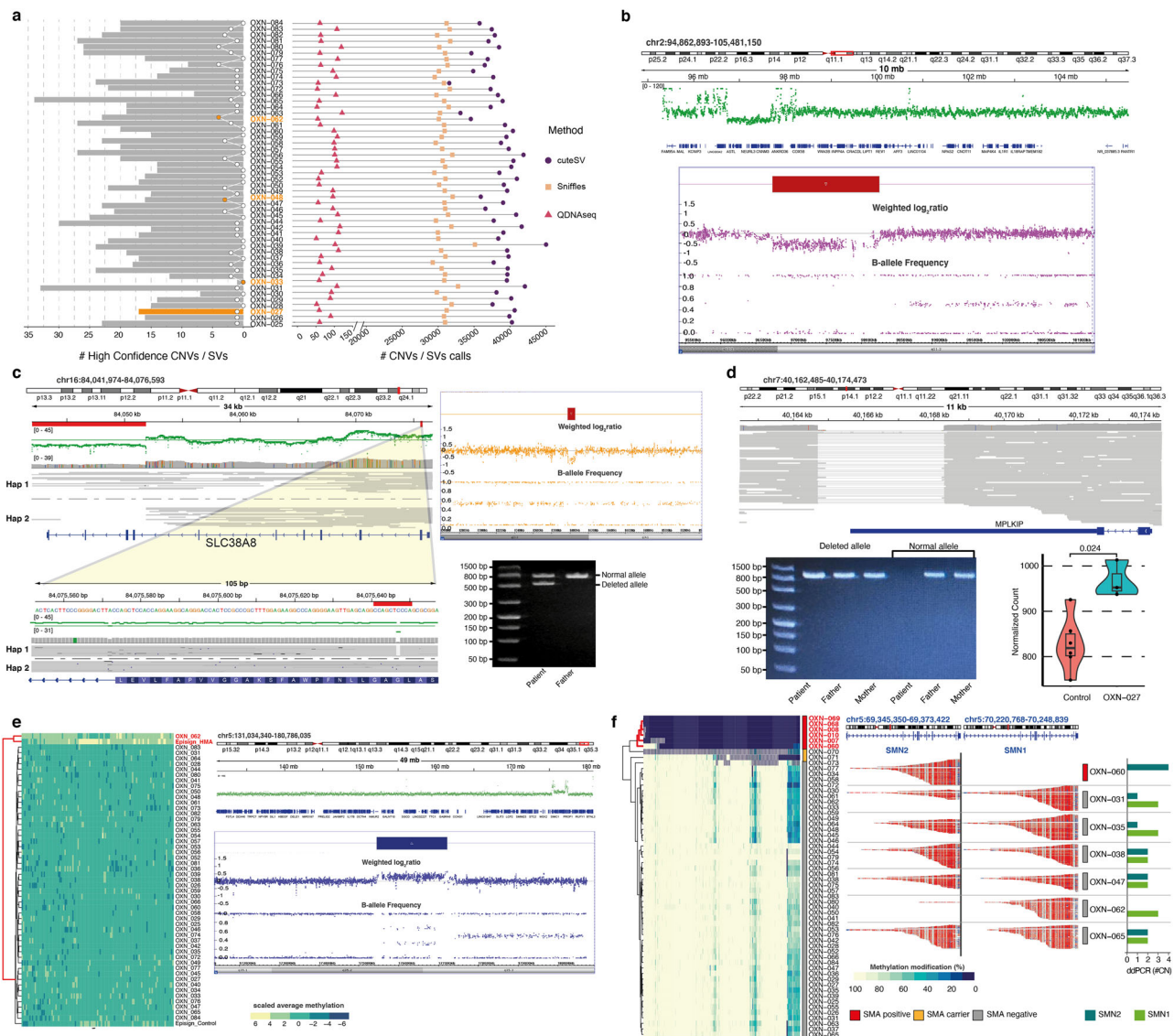


Fig. 2 | Detected pathogenic variants in patients, with previously negative testing, leading to confirmed diagnoses. a Reported on the right side of the graph are the detected numbers of genomic variants (CNVs and SVs), and on the left side, the high confidence variants, post funnel down filtering where samples with confirmed diagnoses, due to CNVs (white line graph) or SVs (gray bars), highlighted in orange. **b** deletion event at 2q11.1-q11.2 in OXN-033 identified by LRS (shown in IGV as a decrease in coverage, in green, in the top panel) and validated by CMA (bottom panel showing log₂ ratios or B-allele frequency (0, 0.5, or 1), in purple, of copy number or SNP probes, respectively. Red bar marks the deletion). **c** Phased genomic alignment with allele-specific INDEL and large deletion in *SLC38A8* in OXN-048 (left top panel) with a zoomed view of INDEL (bottom panel) in IGV. Coverage shown as green dots, and red bars represent intragenic deletion (top) or INDEL (bottom). CMA profile (right top panel showing log₂ ratios or B-allele frequency (0, 0.5, or 1), in orange, of copy number or SNP probes, respectively. Red bar marks the deletion) and PCR gel electrophoresis (bottom panel; refer to Methods) corroborating the finding. **d** homozygous deletion in the 3' UTR of *MPLKIP* detected by LRS (IGV alignment, top panel), validated by PCR (bottom left panel; refer to Methods) with significant difference in the normalized gene expression (bottom

right panel) between Control ($n = 6$; two independent biological samples each repeated 3 times) and OXN-027 ($n = 3$; same sample repeated 3 times) as determined by transcriptomic sequencing (refer to Methods). Box plots lower and upper bounds are 25% and 75% percentile, respectively; the line represents the median, and the whiskers show the minima and maxima of data points. P -value was calculated using the Wilcoxon two-tailed test. **e** left, heatmap with methylation profile across negative cohort (including OXN-62) and a published HMA control sample (left panel); right, duplication event at 5q35.2-q35.3 in OXN-062 detected by LRS (IGV coverage view in green, top right panel) and validated by CMA (bottom right panel showing log₂ ratios or B-allele frequency (0, 0.5, or 1), in blue, of copy number or SNP probes, respectively. Blue bar marks the duplication). **f** heatmap of base methylation modification (%) within the chr5:70239954-70249165 region across the negative cohort, SMA positives (OXN-007, OXN-008, OXN-010, OXN-068, OXN-069) and carriers (OXN-070, OXN-071) (left panel). IGV methylation view post-read deconvolution (center panel) and copy numbers (bar graphs, right panel) of *SMN2* (dark green) and *SMN1* (light green) detected by ddPCR. Source data are provided as a Source Data File. All PCR gels are performed with a 1500 bp ladder, and units are in base pairs (refer to “Methods”).

mechanisms such as positional effects or the disruption of topologically associating domains (TADs). However, such events would still require functional validations which might not yet be part of routine clinical testing. In two patients, for example, transcriptomic analysis was needed to confirm an RNA splicing effect (OXN-044) or to assess an expression outlier (OXN-027).

We also developed an LRS-based “Epimarker” method to empirically profile patients for epismarkers of 36 diseases in the clinical setting. We also uncover, for the first time, an SMA-specific methylation tag which was incorporated into our clinical “Epimarker” profiling. Taken together, our results demonstrate the potential of long-read sequencing as a single unified assay for

routine clinical genetic testing and the discovery of novel rare disease variations.

Methods

Patient samples

This study was reviewed and approved by the Dubai Scientific Research Ethics Committee, Dubai Health Authority (approvals no. DSREC-SR-03/2023_08 and DSREC-08/2024_09). Control DNA samples ($N=17$), with known genomic or methylation aberrations (Supplementary Fig. 1a), were used for optimizing the library preparation, sequencing, bioinformatics analysis, and clinical annotation and filtration. The clinical utility of our approach was then evaluated on DNA from 51 patients with highly suspected monogenic disorders, and non-diagnostic short-read whole exome sequencing. 41% of those patients ($N=21$) had undertaken multiple genetic testing, of which 86% ($N=18$) had received chromosomal microarray (CMA) testing. All patients were consented for clinical genetic testing under an approved de-identified research protocol, which permits the publication of de-identified analyses.

Long read WGS library preparation and sequencing

Genomic DNA was extracted from peripheral whole blood using the QIA-symphony DSP DNA Kit (Qiagen, Hilden, Germany) and QIA-symphony automated nucleic acid extraction instrument, according to the manufacturer's instructions. 6000 ng gDNA was sheared with G-Tubes (Covaris LLC, USA) following the standard 20 kb protocol. The resulting DNA fragments were utilized for duplicate library preparation per sample using the Ligation Sequencing Kit V14 (Oxford Nanopore, UK), according to the manufacturer's instructions. Libraries were sequenced on the PromethION P48 device with R10.4.1 flow cells (Oxford Nanopore, UK) for 72 h with a second library loaded at 24 h post flow cell washing.

mRNA library preparation and Transcriptome sequencing

Transcriptome sequencing was performed for two patients and two controls (Supplementary Data 4). Total RNA was extracted and purified from human whole blood samples collected in Tempus blood RNA tubes using a Tempus spin RNA isolation kit (Applied Biosystems, US), according to the manufacturer's instructions. 270–290 ng of total RNA was utilized for triplicate library preparation per sample using TruSeq® Stranded mRNA Library Prep kit (Illumina, USA), according to the manufacturer's instructions. Libraries were sequenced on Illumina NovaSeq 6000.

Long-read sequencing data analysis

A new pipeline appropriate for long-read nanopore technology was developed in-house using published software (see Supplementary Methods for details). Briefly, base calling was done using “high-accuracy base calling” (HAC) mode during the run using MinKnow distribution (version 22.05.7) and Guppy (version 6.1.5). The methylation tag (MM, ML / mm, ml) was inferred using samtools (version 1.13) for all bam passed files and were aligned to the human reference genome (GRCh37/hg19) using minimap2 (version 2.22-r1101). Epi2Me¹⁸ workflow wf-human-variation (v1.2.0), suitable for long read technology was used for the detection of the genomic variants using its module – ‘-cnv’, ‘-sv’, ‘-snp’ and ‘-methyl’, with default parameters except for CNVs that was run with a bin size of 5. CuteSV (v2.0.3) was applied in conjunction with identifying SVs. CNVs and SVs were annotated using ClassifyCNV (1.1.1) and AnnotSV (v3.2.3). A funnel-down approach was used to filter SVs and CNVs, where SVs with at least 5 supporting reads with allele frequency ≥ 0.3 and CNVs with \log_2 fold change of 0.5 were used for downstream analysis. Variants overlapping coding regions of genes associated with disease as identified from OMIM and GeneCC database were retained, and those unique within the cohort and each method were correlated with patients' phenotype using in-house

scripts. Matching variants were then manually inspected to identify putative pathogenic ones.

Methylation analysis was performed by comparing the methylation profile of the patients with those reported in literature for the epigenomic signature². SMA detection was developed based on the methylation profile in the genomic region capturing introns 6 to 8 (chr5:70,239,954-70,249,165) of *SMN1*, where few bases (0–10%) were observed to undergo methylation modifications indicating absence of *SMN1*.

Single nucleotide variation (SNV) analysis was performed for variants not captured by whole exome sequencing as per ACMG guidelines¹⁴. In addition, variants within 50 bp annotated exon-intron boundary (NCBI Refseq transcripts for build hg19) were analyzed if the splice score as calculated from SpliceAI¹⁹ ≥ 0.7 . Briefly, SNVs with genotype quality, read depth, and mapping quality greater than 10, 30, and 10, respectively with filter tag as “PASS” were selected. Rare variants, detected in ≤ 2 patients in the cohort, present in disease-associated genes as defined by HGMD, OMIM, and GeneCC, where 90% of the isoforms were affected by the variant were selected for manual inspection and correlation with patient phenotype.

Transcriptome sequence data analysis

FastQC and MultiQC were used to assess sequencing read quality. High-quality reads ($Q \geq 30$) were mapped to GRCh37 (hg19) using STAR (v2.7.8a) with the default settings. Gene count was performed using featureCounts from the SubReads (v2.0.1) with the ‘-p -O -g gene_id -s 2’ parameters (Supplementary Data 5) and analyzed by DESeq2 (v1.38.3) correcting for batch effects, normalization and differential gene expression analysis. Genes with adj p -value < 0.05 were identified as significant and selected for pathway enrichment analysis using the Enrichr web application (Supplementary Data 6). Additional statistical analysis was performed using the Fisher exact test to rank the top pathways.

Chromosomal microarray analysis

Chromosomal microarray analysis was performed as previously described²⁰. Briefly, CMA was done using the Affymetrix CytoScan HD™ assay consisting of 2.67 million probes and analyzed using Chromosome Analysis Suite™ software 4.0 to compare, in silico, the hybridization pattern of a patient specimen against a pooled reference sample set. Losses larger than 200 kb (with ≥ 25 probes) or gains larger than 400 kb (≥ 50 probes) are reported, along with smaller variants of pathogenic potential.

Droplet digital PCR analysis

The copy numbers of *SMN1* and *SMN2* were determined by droplet digital PCR (ddPCR) technology as described previously²⁰, using pre-designed proprietary ddPCR assay kits for *SMN1* (Catalog No: 186-3500, Bio-Rad). In addition, experimental controls – 0 copy, 1 copy, and 2 copy controls for *SMN1* were included along with a no template control. Data analysis was performed using QuantaSoft version 1.7.4.0917 (Bio-Rad) to determine the copy number variation (CNV).

PCR Gel electrophoresis

To confirm the heterozygous deletion of approximately 80 kb at 16q23.3 in patient OXN-048 (Fig. 2c), two primer sets of approximately 20 bp in length were designed near the breakpoints (Supplementary Data 7). One set spans the full deletion (chr16:83971418-84052205) with an expected amplification of 500 bp only in carriers of the deleted allele (OXN-048). The other primer set is expected to amplify the 815 bp region from the upstream breakpoint through the deleted region (chr16:83971418-83972232). Gel electrophoresis analysis using a 1500 bp ladder revealed an ~800 bp band for the unaltered allele (Father and OXN-048) and a ~500 bp band for the deleted allele in OXN-048 only (Fig. 2c).

For the validation of a homozygous deletion of 3.6 kb in patient OXN-027 (Fig. 2d), two primer sets were designed (Supplementary Data 7), one spanning the full deletion (chr7:40164307-40168860) but expected to amplify a 952 bp PCR product in carriers of the deleted allele (OXN-27 and parents). The second primer set is expected to amplify a 1045 bp PCR product (chr7:40168214-40168860) spanning the second breakpoint, which cannot be detected in individuals homozygous for the deletion (OXN-027). PCR analysis revealed a ~950 bp band for the deleted allele in the patient and heterozygous parents. However, no band was observed for the patient (OXN-027) using the primer set within the deleted region, indicating a homozygous deletion (Fig. 2d).

All primers were designed using the UCSC Genome Browser, selecting sequences with an optimal length of 20–25 bp, a melting temperature (T_m) of ~60 °C, and a GC content of less than 60%. The final product sizes were confirmed using UCSC In-Silico PCR. PCR products were run on Lonza FlashGel Device (Bioscience).

Statistical analysis

Statistical details of experiments and analyses can be found in the figure legends and the main text. Differences between the two groups were performed using Wilcoxon rank-sum two-tailed test. Comparisons between multiple groups were performed using Wilcoxon pairwise two-tailed test with bonferroni correction. Exact P -values are reported unless smaller than 0.0001.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Raw sequencing data are protected and are not publicly available due to data privacy laws. Processed data from patients who consented for data sharing is publicly available on: <https://figshare.com/s/e58cf877382d23b6b6db>, <https://figshare.com/s/69e52b89dd19c0088134>. Raw data from the independent sample set with known methylation profiles used in validation are available upon request using the European Genome-Phenome Archive (EGA) platform with study number EGAS50000000719²¹. The methylation data from this study used for Epimarker evaluation are available on GitHub: JorisVermeeschLab/NSBEpi (<https://github.com/JorisVermeeschLab/NSBEpi/tree/main>). Source data are provided in this paper.

Code availability

The code developed and used in this study is available on GitHub: https://github.com/Shruti-BioCode/AJCH_ONT_Diagnostic_Utility²².

References

- Kent, A., Parker, A. P., Patel, A., Wynn, S. L. & Steward, C. A. Genomics in rare diseases: an overview for the patient, family and non-specialist healthcare professional. *Future Rare Diseases* **3**, FRD56 (2023).
- Aref-Eshghi, E. et al. Evaluation of DNA methylation epigenatures for diagnosis and phenotype correlations in 42 mendelian neurodevelopmental disorders. *Am. J. Hum. Genet.* **106**, 356–370 (2020).
- Blöß, S. et al. Diagnostic needs for rare diseases and shared pre-diagnostic phenomena: Results of a German-wide expert Delphi survey. *PLoS ONE* **12**, e0172532 (2017).
- Mitsuhashi, S. & Matsumoto, N. Long-read sequencing for rare human genetic diseases. *J. Hum. Genet.* **65**, 11–19 (2020).
- Neerman, N. et al. A clinically validated whole genome pipeline for structural variant detection and analysis. *BMC Genomics* **20**, 545 (2019).
- Oehler, J. B., Wright, H., Stark, Z., Mallett, A. J. & Schmitz, U. The application of long-read sequencing in clinical settings. *Hum Genomics* **17**, 73 (2023).
- Mizuguchi, T. et al. A 12-kb structural variation in progressive myoclonic epilepsy was newly identified by long-read whole-genome sequencing. *J. Hum. Genet.* **64**, 359–368 (2019).
- Miller, D. E. et al. Targeted long-read sequencing identifies missing disease-causing variation. *Am. J. Hum. Genet.* **108**, 1436–1449 (2021).
- Sone, J. et al. Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. *Nat. Genet.* **51**, 1215–1221 (2019).
- Geysens, M. et al. Clinical evaluation of long-read sequencing-based epigenature detection in developmental disorders. *Genome Med.* **17**, 1 (2025).
- Ogino, S. & Wilson, R. B. Genetic testing and risk assessment for spinal muscular atrophy (SMA). *Hum. Genet.* **111**, 477–500 (2002).
- Blasco-Pérez, L. et al. Beyond copy number: A new, rapid, and versatile method for sequencing the entire SMN2 gene in SMA patients. *Hum. Mutat.* **42**, 787 (2021).
- Monani, U. R. et al. A single nucleotide difference that alters splicing patterns distinguishes the SMA gene SMN1 from the copy gene SMN2. *Hum. Mol. Genet.* **8**, 1177–1183 (1999).
- Richards, S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–423 (2015).
- Riley, K. N. et al. Recurrent deletions and duplications of chromosome 2q11.2 and 2q13 are associated with variable outcomes. *Am. J. Med. Genet. Part A* **167**, 2664–2673 (2015).
- Kuht, H. J. et al. SLC38A8 mutations result in arrested retinal development with loss of cone photoreceptor specialization. *Hum. Mol. Genet.* **29**, 2989–3002 (2020).
- Mayr, C. What Are 3' UTRs Doing? *Cold Spring Harb. Perspect. Biol.* **11**, a034728 (2019).
- EPI2ME Labs 23.02-01 Release. EPI2ME Labs <https://labs.epi2me.io/epi2me-labs-23.02.01-release/> (2023).
- Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548 (2019).
- El Naouf, M. et al. The genomic landscape of rare disorders in the Middle East. *Genome Med.* **15**, 5 (2023).
- Clinical evaluation of long read sequencing-based epigenature detection in developmental disorders - EGA European Genome-Phenome Archive. <https://ega-archive.org/studies/EGAS50000000719> (2025).
- Sinha, S. et al. Long read sequencing enhances pathogenic and novel variation discovery in patients with rare diseases. Preprint at <https://doi.org/10.21203/rs.3.rs-4235049/v1> (2025).

Acknowledgements

This work received funding support from Oxford Nanopore Technologies in the form of reagents and consumables. AAT also received funding from the Al Jalila Foundation (grant number MBRU-PD2024-09) and the Mohammed Bin Rashid University of Medicine and Health Sciences at Dubai Health (grant numbers MBRU-RG2024-05 and CSRG-24-19).

Author contributions

A.A.T. conceived the project and obtained funding. F.R. performed long-read DNA and transcriptome sequencing on all samples with support from I.C. M.E.N., R.A., N.H., and S.Y. performed short-read sequencing and other genomic analyses. S.S. developed the code and performed all analysis with support from S.R. R.J., M.S.H., S.Shenbagam, and A.T. helped with data collection and interpretation.

S.Shenbagam and A.T. recruited patients. S.S. and A.A.T. generated the first draft of the manuscript. M.U., M.A.A., S.D.P., and A.A.A. and all coauthors edited the manuscript and provided feedback on the study.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-57695-9>.

Correspondence and requests for materials should be addressed to Shruti Sinha or Ahmad Abou Tayoun.

Peer review information *Nature Communications* thanks Ichizo Nishino, who co-reviewed with Nobuyuki Eura, and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025