

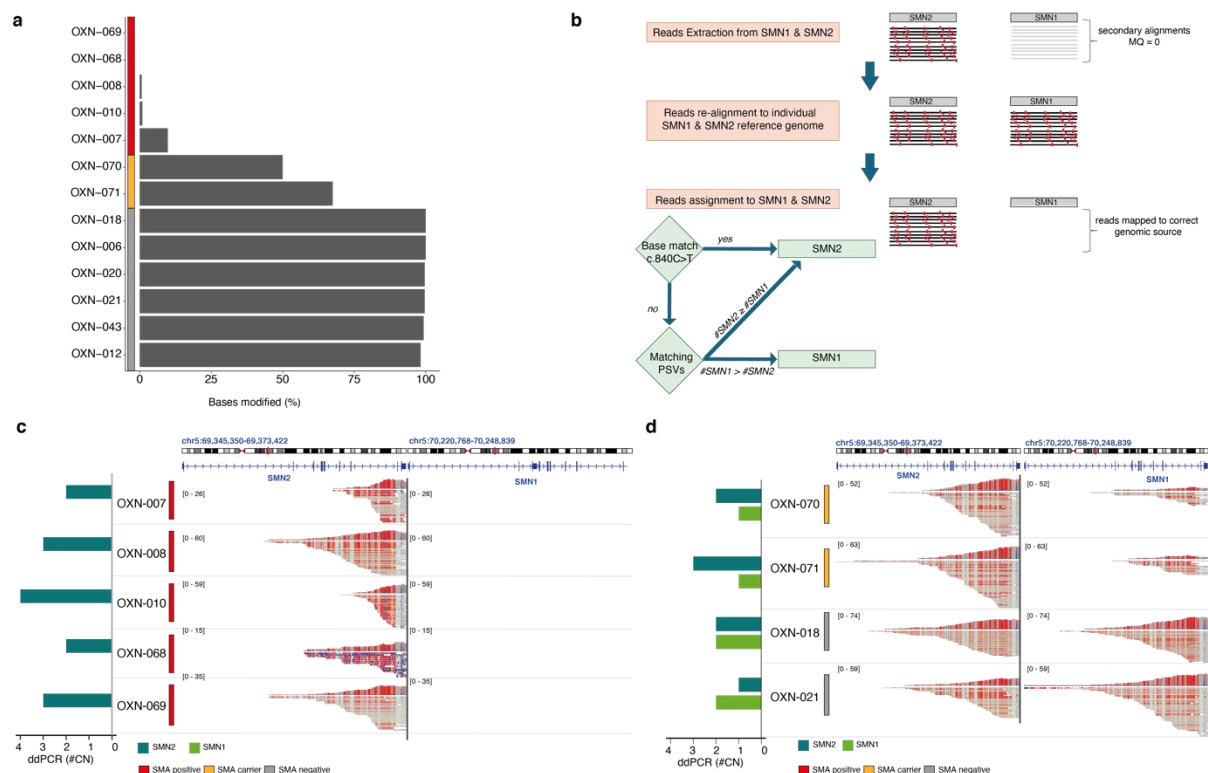
Long read sequencing for pathogenic and novel variation discovery in patients with rare diseases

SUPPLEMENTARY FIGURES



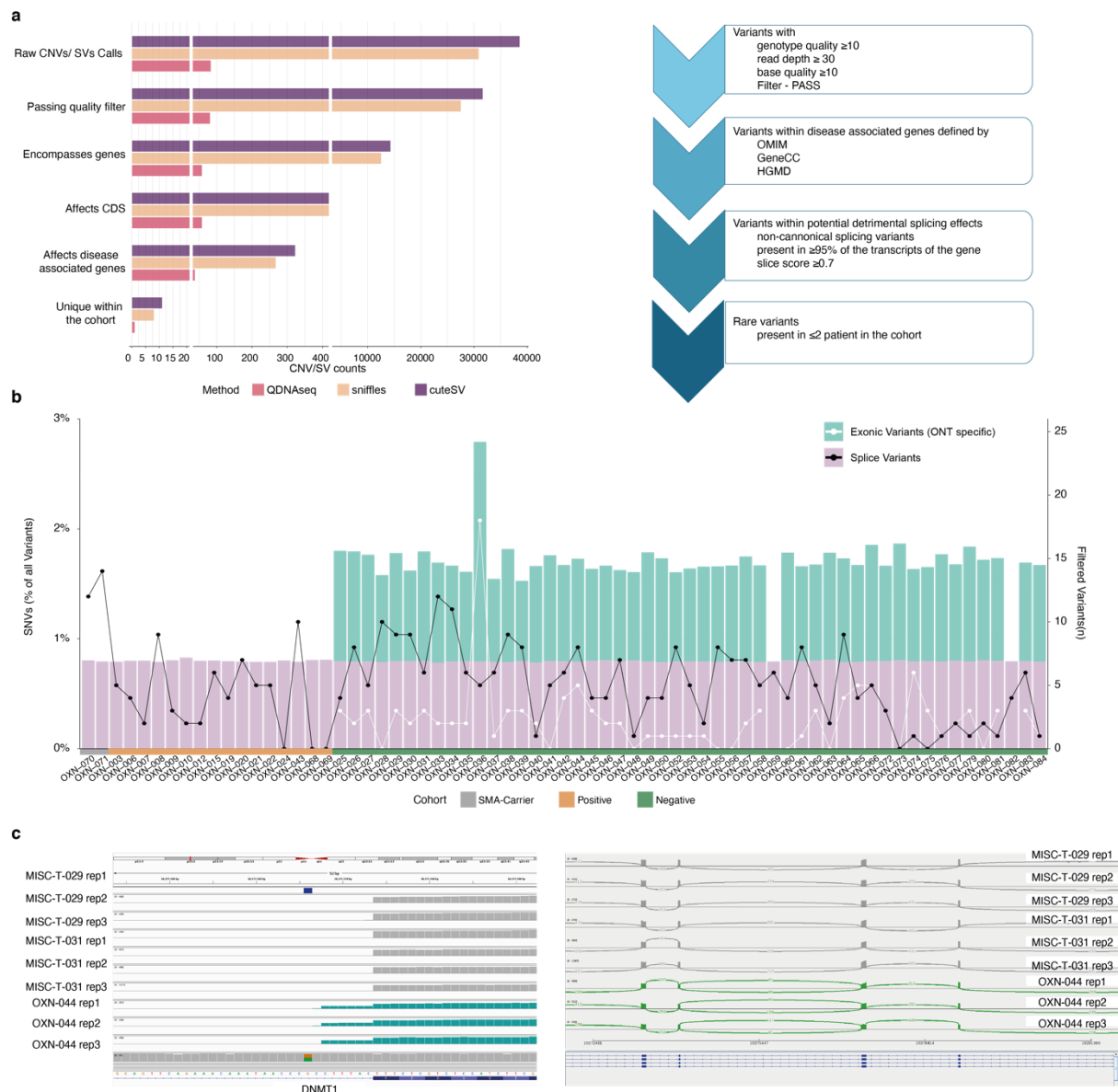
Supplementary Figure 1. Positive cohort description and detected genomic variants. Shown are a) table describing the samples in the positive cohort with the clinical assays conducted confirming the diagnosis. b) Nanopore whole genome sequencing quality showing N50 and coverage. c) Reported on the right side of the graph are the detected number of

genomic variants (CNVs and SVs) and on the left side are high confidence CNVs (white line graph) or SVs (blue bars), post funnel down filtering identified using LRS. d) a complex tandem deletion and duplication event observed in OXN-024 by LRS (top panel, coverage represented in green with solid orange and purple lines on top marking a deletion and a duplication, respectively) and confirmed by CMA (bottom panel showing log₂ ratios or B-allele frequency (0, 0.5, or 1), in blue, of copy number or SNP probes, respectively). Red and blue bar marks the deletion and duplication, respectively) e) heatmap (left) and violin plot (right) with % methylation across all the bases within the Angelman/Prader Willi 15q11.2 CpG 77 region (chr15:25200035-25201054). Statistical comparison by Wilcoxon pairwise two-tailed test with bonferroni correction among OXN-018 (Angelman positive, mean=0.3), OXN-068 (mildly aberrant methylation, mean=15.95) and other negative samples (Angelman negative, mean=48.50). P value<2.2e-16 for all comparisons. f) Shown (top) is the IGV methylation view (top) along with the percentage of heterozygous SNPs across all chromosomes (bottom) of OXN-018 (Angelman positive), OXN-068 (mildly aberrant methylation) and a control. Although we noted mildly reduced methylation in OXN-068 at this locus, it had significantly higher methylation levels compared to the OXN-018 (Angelman positive), with no LOH, (1f). MS-MLPA testing further confirmed that OXN-068 is negative for Angelman Syndrome. Interestingly, OXN-068 is SMA positive and a recent study¹ has linked SMN1 to the Padre-Willi Syndrome/Angelman locus, which we speculate can explain the altered methylation profile at the 15q11.2 locus. Source data are provided as a Source Data file.



Supplementary Figure 2. Methylation profile and read de-convolution for detection and confirmation of Spinal Muscular Atrophy (SMA). Shown are a) Percentage of bases in *SMN1* chr5:70239954-70249165 region observed to have base modification in SMA positive (red), SMA carrier (yellow) and SMA negative (grey) cases b) workflow with schematic visualization

of methylation profile at each step of de-convolution of reads within *SMN1* and *SMN2* homologous region using 16 paralog specific variants (PSVs). The schematic depicts a positive SMA case where no reads are expected to map to *SMN1* after read deconvolution. c) and d) Shown are the IGV visualization of reads aligned to *SMN2* (left) and *SMN1* (right) post de-convolution along with detected copy number by ddPCR (bars) for *SMN2* (dark green) and *SMN1* (light green) in c) SMA positive d) SMA carrier and SMA negative cases. Source data are provided as a Source Data file.



Supplementary Figure 3. Variant statistics on “Funnel down” approach. a) Shown are counts of CNVs and SVs for each method in each filtering step of “funnel down” approach in the negative cohort (left panel) and workflow for filtering SNVs with splicing effects (right panel), b) percentage of total SNVs detected as splicing variants (within 50bp of exon-intron boundary) and long read-specific exonic variants (bars), as well as number of splicing variants (black line) and long read specific exonic variants which are passing filtering criteria and are

present in disease causing genes (white line). c) shown is the effect of a splicing variant in OXN-044 with high spliceAI prediction (>0.7), leading to altered 3' exon with in-frame inclusion of two amino acids as shown by RNAseq data in three replicates (Left, a heterozygous intronic C>A SNV in the *DNMT1* gene leading to a heterozygous extension of the exon by 6 nucleotides, green bars, compared to two independent biological samples, MISC-T-029 and MISC-T-03, each run in triplicates. Right, Sashimi plot showing the introduction of a novel splice junction in OXN-044 (green) compared to controls samples (grey).

SUPPLEMENTARY METHODS

Base Calling And Alignment of Long Read Sequencing Data

Base calling was done using “high-accuracy base calling” (HAC) mode during the run with 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 then aligned to the human reference genome (GRCh37/hg19) using minimap2 (version 2.22-r1101) and the command appropriate for oxford nanopore technology ‘*minimap2 -x map-ont -a -y*’.

Copy Number Variation Analysis of Long Read Sequencing Data

To identify pathogenic copy number variants (CNVs) for definitive diagnoses, a filtering strategy was developed that focused on variants with functional impact on known disease-causing genes. Briefly, CNV calls were generated using QDNAseq² (v1.34.0) available in the cnv module of Epi2Me³ workflow wf-human-variation (v1.2.0) with the bin size of 5 using command “*nextflow run wf-human-variation --cnv --bin_size 5*”. Since, the project aimed to detect pathogenic variants, the focus was placed on CNVs spanning genes known to be associated with diseases. To this end, all CNVs were annotated by ClassifyCNV⁴ against Refseq genes, OMIM⁵, GeneCC⁶ and HGMD and were selected based on the following criteria:

1. Passing quality filter – CNVs with an absolute log2fold change ≥ 0.5 (indicating at least a one-fold change in the copy number) were selected.
2. Encompassing genes – Only CNVs that encompassed genes were selected, excluding variants where the function cannot be explained.
3. Affects coding sequence (CDS) – CNVs affecting coding regions (exon, UTR, transcription start, transcription end sites) that could have functional impact on gene expression were selected.
4. Affects disease associated genes – Only CNVs involving at least one gene known to be associated with disease, were prioritized.
5. Rare CNVs – CNVs containing ≥ 1 disease associated gene not undergoing similar aberration in any other sample within the cohort or with $\leq 30\%$ overlap and ≤ 10 frequency in DGV were prioritized. This strategy help removed common variants present in the cohort. Since homozygous CNVs, indicate functional impact on both the alleles, homozygous CNVs that exhibit similar aberrations in heterozygous condition were categorized as rare CNVs.

Rare CNVs were selected for manual inspection, which included visual validation using IGV and a correlation of the disease-associated gene within the CNV with patient's clinical indication. Only those CNVs that could explain the patient's phenotype were identified as

pathogenic. Given the heterogeneity of our cohort, rare CNVs were selected on per-sample basis. In cases where cohort includes samples with same clinical indications, rare CNVs can be defined per-phenotype group.

Structural Variation Analysis of Long Read Sequencing Data

Structural variations (SVs) calls were made using Sniffles⁷ (v2_2.0.7) available in the 'sv' module of Epi2Me³ workflow wf-human-variation (v1.2.0) and CuteSV⁸(v2.0.3) with default parameters. Since, the study aimed at detection of pathogenic variants for definitive diagnoses, SVs that contained genes known to be associated with diseases were prioritized for the analysis. To achieve this all SVs from the two methods were annotated by AnnotSV⁹ against Refseq genes, OMIM⁵, GeneCC⁶ and HGMD and filtered as per following criteria:

1. Passing quality filter – SVs with ≥ 5 supporting reads and ≥ 0.3 allele frequency of 0.3 were selected.
2. Encompassing genes – Only SVs that encompassed genes were selected, excluding variants where the function cannot be explained.
3. Affects coding sequence (CDS) – SVs affecting coding regions (exon, UTR, transcription start, transcription end sites) that could have functional impact on gene expression were selected.
4. Affects disease associated genes – SVs involving at least one gene known to be associated with disease, were prioritized.
5. Rare SVs - SVs containing ≥ 1 disease associated gene not undergoing similar aberration in any other sample within the cohort or with $\leq 30\%$ overlap and ≤ 10 frequency in DGV were prioritized. Since homozygous SVs, indicate functional impact on both the alleles, homozygous SVs that exhibit similar aberrations in heterozygous condition were categorized as rare SVs.

Rare SVs were prioritized for manual inspection which included visual validation using IGV and a correlation of the disease-associated gene within the SVs with patient's clinical indication. Only those SVs that could explain the patient's phenotype were identified as pathogenic. Given the heterogeneity of our cohort, rare SVs were selected on per-sample basis. In cases where cohort includes samples with same clinical indications, rare SVs can be defined on per-phenotype group.

Splicing Analysis of Small Nucleotide Variants in Long Read Sequencing Data

Short nucleotide variations were detected using Claire3 in phase mode available in the 'snp' module of Epi2Me³ workflow wf-human-variation (v1.2.0). Variants within 50bp of the coding region boundaries were assessed for their splicing effect using SpliceAI¹⁰. Variants were selected for manual inspection if they passed all filtering criteria.

1. Variants passing quality score - Variants with genotype quality ≥ 10 , read depth ≥ 30 , base quality ≥ 10 , filter tag as "PASS"
2. Variants within disease associated genes – Variants were annotated for genes based on NCBI Refseq transcripts for build hg19 and those within genes associated with a disease as defined in HGMD, OMIM⁵ and GeneCC⁶ database were selected.
3. Variant with potential splicing effect – Variants with splice score of ≥ 0.7 and present in majority of the isoforms (90%) were considered as high confidence splicing variants.
4. Rare variants – Variants present in ≤ 2 samples in the cohort

Long Read Sequencing-Specific Small Nucleotide Variant Analysis

Short nucleotide variations were detected using Claire3 in phase mode available in the 'sv' module of Epi2Me³ workflow wf-human-variation (v1.2.0). Variants detected previously by genetic test by whole exome sequencing or not in exons were removed from further analysis. The remaining variants were annotated with RefSeq, OMIM, ClinVar and HGMD by VarSeq¹¹ (version: 2.6.1). Variants meeting all the following criteria were selected for manual inspection if the genes could explain patient's phenotype.

1. Variants passing quality score - Variants with genotype quality ≥ 10 , read depth ≥ 30 , base quality ≥ 10 , filter tag as "PASS".
2. Variants with allele frequency ≥ 0.16 classified as disease causing mutation (DM), pathogenic, likely pathogenic or with loss of function (disruptive_inframe_insertion, disruptive_inframe_deletion, exon_loss_variant, frameshift_variant, inframe_insertion, inframe_deletion, stop_lost, stop_retained_variant, stop_gained) as well as splicing variants within ± 2 bp of exon boundary with splice score ≥ 0.5
3. Rare variants – Variants present in ≤ 2 samples in the cohort with homozygous exome count ≤ 2 and total filtered homozygous count ≤ 2

Epimarker - EPISIGN Methylation Profile Matching Analysis

Methylation analysis was performed by comparing the methylation profile of the patients with those reported in literature for the epigenomic signature¹² for a total of 34 Mendelian neurodevelopment disorders (MNDD). For the purpose of comparison; disease specific probes from Illumina Infinium methylation 450k and EPIC bead chip arrays identified as episignature in the paper were mapped on the human genome hg19 using pblat¹³ the parameter “-fastMap”. In order to remove ambiguity coming from multimapping probes; probes with block count of 1 with alignment length matching the probe length were selected and assessed for the downstream analysis. Aggregated methylation modification counts for each base from long read sequencing in the probe region were estimated using modbam2bed from 'methyl' modules of Epi2Me³ workflow wf-human-variation (v1.2.0). Methylation values for all sample and MNDD were standardized (i) and normalized (ii) using min-max normalization using the equation below where s is the sample, p is the probe, x_p is the methylation value for each probe, $\bar{x}$ is the mean and σ is the standard deviation, $stdMethyl$ is the standardized methylation value and $normMethyl$ is normalized methylation value

$$(i) \quad stdMethyl_p = \frac{\bar{x}_s - x_p}{\sigma_s}$$

$$(ii) \quad normMethyl_p = \frac{stdMethyl_p - \min(stdMethyl_s)}{\max(stdMethyl_s) - \min(stdMethyl_s)}$$

Since, there is limitation of the availability of the training datasets used for the MNDD EpiSigns, and there is no universal clustering and distance matrix that is fit for all conditions and data^{14,15}, we used multiple matrices and clustering algorithm for assignment of the sample methylation profile to a MNDD class. Briefly, hierarchical clustering was performed using euclidean distance and ward.D2 for MNDD epi-signature probes on normalized methylation values for each sample with the MNDD. Euclidean distance, principal component distance, mutual information and correlation matrix was computed for the normalized methylation

modification. For a MNDD episignature probes, sample and MNDD were assigned cluster group as below .

1. Nearest Neighbour : Reciprocal nearest neighbour with the MNDD was observed based on euclidian or/and PCA distance from normalized methylation.
2. Hierarchical Group: Sample clustered with the MNDD in hierarchical clustering.
3. Highest correlation: MNDD with highest normalized methylation correlation where MNDD is the nearest neighbour for the sample based on correlation distance for the respective episignature probes.
4. Highest mutual information: MNDD with the highest mutual information among MNDDs that are the nearest neighbour for the sample based on mutual information for the respective episignature probes.

For each sample, the association with the MNDD was evaluated using the Fisher's test, based on how often a sample was the nearest neighbor to a MNDD across the 34 MNDD episignature probes. The matrix values for PCA, Euclidean distance, and mutual information were used for this analysis. P-values were adjusted for multiple comparisons within each sample using the p.adjust function in R to control the False Discovery Rate (FDR). A sample was classified as a MNDD if the normalized methylation correlation with the MNDD > 0.7 and it met one of the following criteria: sample and MNDD clustered in one cluster group with an adjusted p-value < 0.001, or if sample and MNDD clustered in more than one cluster group with an adjusted p-value < 0.5. All samples not classified to a MNDD was assigned as "non-MNDD".

An independent dataset of methylation modification from long reads, comprising of a total of 60 samples¹⁶ was used for Epimarker evaluation. Samples "EPI_13", "EPI_18", "EPI_20" were removed from the validation set, as clinical tests did not confirm the methylation signature for these samples. Overall, the dataset included 9 MNDD with 17 disease samples and 40 controls. Epimarker showed 100% sensitivity and 100% specificity with all disease samples correctly classified and none of the control samples assigned to a MNDD (Supplementary Data 7). Cornelia de Lange (CdLS), Autism susceptibility to 18 (AUTS18/CHD8_1 and AUTS18/CHD8_8) and Rubinstein-Taybi syndrome (RSTS/CREBBP) have been reported to exhibit poor reproducibility in methylation signature¹⁷. While Epimarker correctly classified these disease samples in the validation dataset, to avoid potentially false diagnosis arising from weak or overlapping methylation profiles reported in these disorders^{12,17}, they were excluded from the MNDD profile set for the analysis by Epimarker.

Spinal Muscular Atrophy (SMA) Detection and Confirmation using Aggregated Methylation Profile and Read De-convolution.

SMN1 genomic region was analysed for the methylation pattern, with aggregated methylation modification counts for each base (≥ 5 supporting reads and $\geq 98\%$ score) using modbam2bed from 'methyl' modules of Epi2Me³ workflow wf-human-variation (v1.2.0). The methylation modification pattern was then compared across all the samples within the positive cohort. The genomics region chr5:70,239,954-70,249165 was identified as specific to SMA, with 0-10% (low), 50-70% (moderate) and 98-100% (high) of bases showing methylation modification in SMA patients, carriers and non-carriers, respectively. Given the homology between *SMN1* and *SMN2*, the observed methylation profile likely results from allelic loss of *SMN1*, seen as secondary alignment with zero mapping quality. To address this, reads were, re-assigned to either *SMN2* or *SMN1* based on 16 paralogous sequence variants (PSVs)^{18,19} (Supplementary Fig S2). Reads spanning the PSV regions of *SMN1*(chr5:70245776-70248601) and *SMN2*

(chr5:69370351-69373181) were re-aligned individually to the respective reference sequence of *SMN1* and *SMN2*. Reads were then reassigned to *SMN1/SMN2* as per the below criteria:

1. PSV c840>T (g.chr5:70247773C>T) modification has been identified to be a gene conversion event from *SMN1* to *SMN2*, hence reads with base modification from g.chr5:70247773C>T were assigned to *SMN2*.
2. Reads were assigned to *SMN1/SMN2* based on highest number of matching PSVs.
 - a. *SMN1* if #PSVs (*SMN1*) > #PSVs (*SMN2*).
 - b. *SMN2* if #PSVs (*SMN2*) > #PSVs (*SMN1*)

In order to assess the occurrence of SMA within the negative cohort, methylation modifications were assessed and reassignment of reads was performed across all samples. Samples were considered as SMA positive, if

1. $\leq 15\%$ of the bases in the region underwent methylation modification.
2. No reads were observed in *SMN1* post re-assignment

Angelman Methylation Analysis

In order to assess the samples for Angelman syndrome, known genomic region 15q11.2 (CpG 77, chr15:25200035-25201054) used for diagnostic purpose in MS-MLPA genetic testing²⁰ was scanned for methylation modification. Samples with $\geq 80\%$ of the bases in the region not undergoing methylation modification with uniparental disomy ($\geq 80\%$ homozygous variants) as estimated by allele frequency were considered as positive for Angelman Syndrome.

Transcriptome Data Analysis

FastQC²¹ and MultiQC²² were used to assess sequencing read quality. High-quality reads ($Q \geq 30$) were mapped to GRCh37 (hg19) using STAR²³ (v2.7.8a) with the default settings. Gene count was performed using featureCounts²⁴ from the SubReads (v2.0.1) with the '-p -O -g gene_id -s 2' parameters. DESeq2²⁵ (v1.38.3) was used for assessing batch effects, normalization and gene expression analysis. Genes with adj p-value < 0.05 were identified as significant and selected for downstream analysis. Enrichr web-application was used for pathway enrichment analysis to identify enriched pathways among differentially expressed genes. All of the differentially expressed genes in each comparison were utilized as input, which was then compared to gene-set libraries in Enrichr. Additional statistical analysis was performed using the Fisher exact test to rank the top pathways.

Droplet Digital PCR Analysis

The copy numbers of *SMN1* and *SMN2* were determined by droplet digital PCR (ddPCR) technology as described previously²⁶, using predesigned proprietary ddPCR assay kits for *SMN1* (Catalog No: 186-3500, Bio-Rad) and *SMN2* (Catalog No: 186-3503, Bio-Rad). The predesigned Bio-Rad assays include reagents to detect both the target gene (*SMN1* or *SMN2*) and an internal reference gene (*RPP30*). In addition, experimental controls – 0 copy, 1 copy and 2 copy controls for *SMN1*, and 2 copy, 3 copy and 4 copy controls for *SMN2* – 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).

Chromosomal Microarray Analysis

Chromosomal microarray analysis was performed as previously described²⁶. Briefly CMA was done using the Affymetrix CytoScan HD™ assay, a high density, whole genome microarray with 2.67 million probes. The array consists of 1.95 million non-polymorphic markers and around 750,000 single-nucleotide polymorphism (SNP) markers, with an average genome-wide probe spacing of 1.1 kb, and an average intragenic probe spacing of 880bp. The assay covers (25 markers/100Kb) 100% ClinGen (formerly ISCA) constitutional genes (n=3,483), 100% OMIM morbid genes (n=3,561) and 98% RefSeq genes (n=36,121). This assay utilizes the 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.

REFERENCES

1. Heimdörfer, D. *et al.* Truncated variants of MAGEL2 are involved in the etiologies of the Schaaf-Yang and Prader-Willi syndromes. *The American Journal of Human Genetics* **111**, 1383–1404 (2024).
2. Scheinin, I. *et al.* DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. *Genome Res.* **24**, 2022–2032 (2014).
3. EPI2ME Labs 23.02-01 Release. *EPI2ME Labs* <https://labs.epi2me.io/epi2me-labs-23.02.01-release/> (2023).
4. Gurbich, T. A. & Ilinsky, V. V. ClassifyCNV: a tool for clinical annotation of copy-number variants. *Sci Rep* **10**, 20375 (2020).
5. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research* **43**, D789–D798 (2015).
6. DiStefano, M. T. *et al.* The Gene Curation Coalition: A global effort to harmonize gene–disease evidence resources. *Genetics in Medicine* **24**, 1732–1742 (2022).
7. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* **15**, 461–468 (2018).
8. Jiang, T. *et al.* Long-read-based human genomic structural variation detection with cuteSV. *Genome Biology* **21**, 189 (2020).
9. Geoffroy, V. *et al.* AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics* **34**, 3572–3574 (2018).
10. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548.e24 (2019).
11. VarSeq | Golden Helix, Inc. <https://www.goldenhelix.com/products/VarSeq/>.

12. Aref-Eshghi, E. *et al.* Evaluation of DNA Methylation Episignatures for Diagnosis and Phenotype Correlations in 42 Mendelian Neurodevelopmental Disorders. *The American Journal of Human Genetics* **106**, 356–370 (2020).
13. Wang, M. & Kong, L. pblat: a multithread blat algorithm speeding up aligning sequences to genomes. *BMC Bioinformatics* **20**, 28 (2019).
14. Oyewole, G. J. & Thopil, G. A. Data clustering: application and trends. *Artif Intell Rev* **56**, 6439–6475 (2023).
15. Jaskowiak, P. A., Campello, R. J. & Costa, I. G. On the selection of appropriate distances for gene expression data clustering. *BMC Bioinformatics* **15**, S2 (2014).
16. Geysens, M. *et al.* Clinical evaluation of long-read sequencing-based episignature detection in developmental disorders. *Genome Med* **17**, 1 (2025).
17. Husson, T. *et al.* Episignatures in practice: independent evaluation of published episignatures for the molecular diagnostics of ten neurodevelopmental disorders. *Eur J Hum Genet* **32**, 190–199 (2024).
18. Monani, U. R. *et al.* A Single Nucleotide Difference That Alters Splicing Patterns Distinguishes the SMA Gene SMN1 From the Copy Gene SMN2. *Human Molecular Genetics* **8**, 1177–1183 (1999).
19. Blasco-Pérez, L. *et al.* Beyond copy number: A new, rapid, and versatile method for sequencing the entire SMN2 gene in SMA patients. *Human Mutation* **42**, 787 (2021).
20. Henkhaus, R. S. *et al.* Methylation-Specific Multiplex Ligation-Dependent Probe Amplification and Identification of Deletion Genetic Subtypes in Prader-Willi Syndrome. *Genet Test Mol Biomarkers* **16**, 178–186 (2012).
21. Andrews, S. FastQC A Quality Control tool for High Throughput Sequence Data. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.

22. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
23. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
24. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
25. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).
26. El Naofal, M. *et al.* The genomic landscape of rare disorders in the Middle East. *Genome Medicine* **15**, 5 (2023).