Clinical Epigenetics

RESEARCH

Open Access

Diagnostic utility of single-locus DNA methylation mark in Sotos syndrome developed by nanopore sequencing-based episignature

Takeshi Mizuguchi[1*], Nobuhiko Okamoto[2], Taiki Hara[1], Naoto Nishimura[1], Masamune Sakamoto[1], Li Fu[1], Yuri Uchiyama[1,3], Naomi Tsuchida[1,3], Kohei Hamanaka[1], Eriko Koshimizu[1], Atsushi Fujita[1], Kazuharu Misawa[1], Kazuhiko Nakabayashi[4], Satoko Miyatake[1,5] and Naomichi Matsumoto[1,3,5*]

Abstract

Background In various neurodevelopmental disorders (NDDs), sets of differential methylation marks (referred to as DNA methylation signatures or episignatures) are syndrome-specific and useful in evaluating the pathogenicity of detected genetic variants. These signatures have generally been tested using methylation arrays, requiring additional experimental and evaluation costs. As an alternative, long-read sequencing can simultaneously and accurately evaluate genetic and epigenetic changes. In addition, genome-wide DNA methylation profiling with more complete sets of CpG using long-read sequencing (than methylation arrays) may provide alternative but more comprehensive DNA methylation signatures, which have yet to be adequately investigated.

Methods Nine and seven cases of molecularly diagnosed Sotos syndrome and ATR-X syndrome, respectively, were sequenced using nanopore long-read sequencing, together with 22 controls. Genome-wide differential DNA methylation analysis was performed. Among these differential DNA methylation sites, a single-locus DNA methylation mark at part of the *NSD1* CpG island (CpGi) was subsequently studied in an additional 22 cases with a *NSD1* point mutation or a 5q35 submicroscopic deletion involving *NSD1*. To investigate the potential utility of a single-locus DNA methylation test at *NSD1* CpGi for differential diagnosis, nine cases with *NSD1*-negative clinically overlapping overgrowth intellectual disability syndromes (OGIDs) were also tested.

Results Long-read sequencing enabled the successful extraction of two sets of differential methylation marks unique to each of Sotos syndrome and ATR-X syndrome, referred to as long-read-based DNA methylation signatures (LR-DNAm signatures), as alternatives to reported DNA methylation signatures (obtained by methylation array). Additionally, we found that a part, but not all, of the *NSD1* CpGi were hypomethylated compared with the level in controls in both cases harboring *NSD1* point mutations and those with a 5q35 submicroscopic deletion. This difference in methylation is specific to Sotos syndrome and lacking in other OGIDs.

Conclusions Simultaneous evaluation of genetic and epigenetic alterations using long-read sequencing may improve the discovery of DNA methylation signatures, which may in turn increase the diagnostic yields. As

*Correspondence:
Takeshi Mizuguchi
tmizu@yokohama-cu.ac.jp
Naomichi Matsumoto
naomat@yokohama-cu.ac.jp
Full list of author information is available at the end of the article

an example of the outcomes of these analyses, we propose that a single-locus DNA methylation test at *NSD1* CpGi may streamline the molecular diagnosis of Sotos syndrome, regardless of the type of *NSD1* aberration.

Keywords DNA methylation signature, Long-read sequencing, Sotos syndrome, ATR-X syndrome, *NSD1*, *ATRX*, Variant of uncertain significance, Nanopore, PacBio, Adaptive targeted long-read sequencing

Background

DNA methylation is a fundamental epigenetic mark regulating gene expression and playing important roles in many biological processes, such as genomic imprinting and X chromosome inactivation, as well as disease pathophysiology [1, 2]. Genes associated with epigenetic regulation have been identified, with their aberrations causing neurodevelopmental disorders (NDDs) and occasionally leading to syndrome-specific differential methylation patterns, possibly through the disruption of epigenetic regulation [3, 4]. Methylation profiling technologies, such as methylation array, enable the identification of sets of differential methylation marks in certain Mendelian NDDs [5–8]. These global and multi-locus DNA methylation patterns in diseases, referred to as DNA methylation signatures (DNAm signatures), may or may not be related to the primary disease pathophysiology.

In the era of next-generation short-read sequencing (NGS), there are many genetic alterations whose impact on health and disease conditions remains unclear, which are defined as variants of uncertain significance (VUS) and present a challenge for achieving a genetic diagnosis. With the aid of disease-specific DNAm signatures, some VUS can be classified as "pathogenic" or "benign." More than 30 diseases are now known to have specific DNAm signatures [4], and this diagnostic utility has led to these signatures being proposed as biomarkers [9].

The recently developed technology of long-read genome sequencing (LR-GS) allows the evaluation of more complete sets of genetic alterations, including single-nucleotide variants (SNVs), structural variants (SVs), complex SVs, repeat expansion mutations, and copy number variations (CNVs), even at repetitive regions at which sequencing is technically challenging using the earlier technologies [10]. In addition to the detection of genetic variants, this technology can simultaneously detect DNA methylation [11, 12]. In fact, several examples of disease-relevant repeat expansion mutations with altered DNA methylation have been discovered from a single long-read sequencing dataset [13–15]. As such, parallel and simultaneous evaluation of genetic and epigenetic alterations, including DNAm signatures, may improve the accuracy of genetic tests. Hence, LR-GS is a promising one-stop genetic and epigenetic test to identify pathogenic variants. To determine the feasibility of this approach, the utility of DNA methylation analysis using long-read sequencing for diagnosing Sotos syndrome is here studied and discussed.

Methods

Nanopore long-read genome sequencing (nanopore LR-GS)

Genomic DNA was extracted from peripheral whole blood. The extracted DNA was initially evaluated for its quality and quantity by Qubit BR (Invitrogen) and pulsed-field gel electrophoresis using CHEF Mapper XA (Bio-Rad), respectively. Basically, a short-read eliminator kit (PacBio) was used to deplete short DNA fragments and enhance the sequencing of long DNA fragments. Depending on their DNA size distribution, samples with extremely high molecular weight (> 100 kb) were fragmented using a Megarupter 2 (Diagenode) with long hydropores (Diagenode) with the size distribution of targeted fragments set to "50 kb" to increase the sequencing throughput. Three micrograms of DNA was subjected to library preparation using the SQK-LSK109 or SQK-LSK110 kit (Oxford Nanopore Technologies), in accordance with the manufacturer's instructions. The prepared libraries were subdivided into two or three aliquots of 20–50 fmol. The first aliquot was loaded onto the FLO-PRO002 flow cell (R9.4.1) and sequenced on the PromethION sequencer (Oxford Nanopore Technologies). To maximize the sequencing throughput of each sample from a single flow cell, the loaded libraries were then digested and removed using the flow cell wash kit EXP-WSH004 (Oxford Nanopore Technologies) to recover the viable pores for sequencing during a 72-h sequencing run. The second and/or third aliquots were reloaded to the washed flow cell. A single PromethION flow cell was used for each sample (a single flow cell per sample), except for in the cases of Ctr1 and Ctr2 (two flow cells per sample).

We generated LR-GS data for 3 Sotos syndrome (SoS1 to SoS3), 3 ATR-X syndrome (ATRX-1 to ATRX-3) and 8 healthy controls (Ctr1 to Ctr8) as a discovery cohort. For validation cohort, we generated LR-GS data for one Sotos syndrome (SoS4), one ATR-X syndrome (ATRX-4), 8 healthy controls (Ctr9 to Ctr16) and 6 *NSD1*- and *ATRX*-negative NDDs (NDD-1 to NDD-6). Sample information is listed in Supplementary Table 1.

Nanopore adaptive targeted long-read sequencing (adaptive T-LRS)

Sequence libraries for adaptive T-LRS were prepared as for nanopore LR-GS. DNA was fragmented using a Megarupter 2 with the size distribution of targeted fragments set to "40 kb" or "20 kb" (Diagenode) to achieve efficient enrichment of the target sequence. The sequence library was loaded onto the MinION FLO-MIN106D Flow Cell (R9.4.1) with a GridION device (Oxford Nanopore Technologies). Sequencing and live basecalling in the high-accuracy basecalling (HAC) model with the adaptive sampling enrichment experiment option were set up in MinKNOW (version 21.05.8 or 21.11.7). Real-time mapping and enrichment of the reads with regions of interest (ROI) were performed against the reference human genome GRCh38 alongside a bed file with the genomic coordinates of the target region. After the first flow cell wash, another sequence library from a different sample was loaded to allow multi-sample sequencing on the same flow cell (a single flow cell for two samples), except for in the case of SoS5 (a single flow cell per sample). We generated adaptive T-LRS data for 5 Sotos syndrome (SoS5 to SoS9), 3 ATR-X syndrome (ATRX-5 to ATRX-7) as a validation cohort. Sample information is listed in Supplementary Table 1.

Nanopore data analysis

We used Guppy basecaller (v6.5.7) to detect 5-methylcytosine (5-mC) using fast5 files as an input. The specific basecalling model for modified bases was processed with the configuration file named "dna_r9.4.1_450bps_modbases_5mc_cg_hac_prom.cfg." Unaligned bam files with 5-mC (MM and ML tag) information from guppy basecaller were converted to fastq using samtools fastq command with the -T 1 option to retain 5-mC information. Reads were mapped to the human reference genome GRCh38 using minimap2 (v.2.26) with -x map-ont [16]. Then, small variant calling and phasing the small variants were performed using the pepper-margin-deepvariant pipeline (v.0.8) with the --ont_r9_guppy5_sup option, and the reads were phased (i.e., an HP tag was added to bam files) [17]. As a result, bam files with both 5-mC (MM and ML tag) and haplotype (HP tag) information were obtained. To generate summary counts of modified (5-mC) and unmodified bases from aligned bam files with MM, ML and HP tags, we used the Modkit program (https://github.com/nanoporetech/modkit) and created bedMethyl format tables for each sample using the following command:

modkit pileup --cpg --ref <reference.fasta> --ignore h --combine-strands

To construct a separate bedMethyl format table of each haplotype (haplotype 1 and haplotype 2), modkit pileup was run with the --partition-tag HP option using the following command:

modkit pileup --cpg --ref <reference.fasta> --ignore h --combine-strands --partition-tag HP.

PacBio high-fidelity long-read genome sequencing (HiFi LR-GS)

Three micrograms of genomic DNA were fragmented using a Megarupter 3 (Diagenode) with the shearing speed setting of 29–31, which depends on the DNA size distribution of each sample. The sheared genomic DNA was purified using 1 × SMRTbell cleanup beads (PacBio, 102-158-300), and the size distribution was checked using a Femto pulse capillary electrophoresis system (Agilent). Two to three micrograms of sheared DNA was subjected to library preparation using an SMRTbell Prep Kit 3.0 (PacBio, 102-182-700), in accordance with the manufacturer's instructions (procedure and checklist-preparing whole-genome and metagenome libraries using SMRTbell prep kit 3.0). The resulting SMRTbell DNA library was size-selected using diluted (35% v/v) AMPure PB beads to deplete short DNA fragments (PacBio, 102-182-500). After completing the AMPure PB bead size selection, the SMRTbell library was annealed with sequencing primers at 23 °C for 15 min, and then primer-annealed SMRTbell template DNA was incubated with sequencing polymerase from the Revio Polymerase Kit at 23 °C for 15 min (PacBio, 102-817-600). Polymerase-bound SMRTbell complex was then purified using SMRTbell cleanup beads (PacBio, 102-817-600). The purified complex was then loaded onto the RevioSMRT Cell (PacBio, 102-202-200) with an on-plate loading concentration of 225 pM, following the instructions of the SMRTlink sample setup module. Samples were sequenced on the PacBio Revio system using a Revio sequencing plate (PacBio, 102-587-400), and data were collected for 30 h using a single flow cell per sample. We generated PacBio HiFi LR-GS data for the 3 members from one trio (SoS10 and his father and mother). Sample information is listed in Supplementary Table 1.

PacBio HiFi data analysis

All datasets were processed in the same fashion using Revio system v13 on-instrument analysis for basecalling, consensus calling (CCS: circular consensus sequencing analysis), and methylation calling (5-mC analysis). Automatically generated HiFi reads bam (> 99% accuracy, > Q20) with 5mC calls at CpG sites (MM and ML tags) from Revio on-instrument analysis were used for downstream analysis. Read alignment, variant calling, phasing, and 5-mC analysis were performed on the

PacBio WGS Variant Pipeline HiFi-human-WGS-WDL (https://github.com/PacificBiosciences/HiFi-human-WGS-WDL). Reads were aligned to the GRCh38 human reference genome using pbmm2 (v1.10.0) with --preset HIFI. Small variant calling and structural variant calling were performed using deepvariant (v.1.5.0) and pbsv (v.2.9.0), respectively. HiPhase (v1.0.0) jointly phased small variants and structural variants and phased the reads (i.e., created bam files with MM, ML, and HP tags) [18]. For downstream methylation analysis using PacBio datasets, the site methylation probabilities from mapped HiFi reads were calculated and reported as bed file outputs using pb-CpG-tools (v2.3.2) with the default options (--model, --denovo) (https://github.com/PacificBiosciences/pb-CpG-tools).

Differential methylation analysis with nanopore sequencing

Differentially methylated regions (DMRs) were called using the DSS (Dispersion Shrinkage for Sequencing data) program [19] with the default parameters, except for minCG = 5 instead of minCG = 3. Inputs for DSS were extracted from bedMethyl format tables, including chromosome number, genomic coordinates, total number of reads, and number of reads showing methylation for each CpG position. DMRs detected by DSS were annotated using the Homer annotate Peaks.pl program (http://homer.ucsd.edu/homer/ngs/annotation.html). We filtered out CpG sites with extremely low or high coverage (i.e., excluding those with < 5× coverage or coverage exceeding three times the mean of each dataset).

Construction of prediction model

The "ksvm" function with type = "C-svc," kernel = "vanilladot," and prob.model = "TRUE" from the kernlab R package was used for a support vector machine (SVM) for classification using DNA methylation data from Sotos syndrome ($n = 3$, SoS1 to SoS3), ATR-X syndrome ($n = 3$, ATRX-1 to ATRX-3), and healthy controls ($n = 8$, Ctr1 to Ctr8). The "predict" function under the SVM model that was trained using the discovery cohort was used with function type = "probabilities" to obtain the probability score for each class (i.e., probability scores for Sotos syndrome, ATR-X syndrome, and healthy controls) in the validation cohort [Sotos syndrome ($n = 6$, SoS4 to SoS9), ATR-X syndrome ($n = 4$, ATRX-4 to ATRX-7), healthy controls ($n = 8$, Ctr9 to Ctr16), and *NSD1*- and *ATRX*-negative NDDs ($n = 6$, NDD-1 to NDD-6)]. Individuals with scores above 0.5 were classified as having a Sotos or ATR-X-associated methylation profile, as previously described [4].

Multi-dimensional scaling (MDS)

To visualize the level of similarity of methylation profiles of Sotos syndrome and ATR-X syndrome, we performed multi-dimensional scaling using the cmdscale function of R, version 3.6.2.

Methylation array

Genome-wide methylation was measured using the Illumina Infinium MethylationEPIC BeadChip array, as described previously [20]. Methylation β values processed by GenomeStudio software were used for the comparison with nanopore methylation analysis. We generated methylation array data for 2 Sotos syndrome (SoS1 and SoS2), 2 ATR-X syndrome (ATRX-1 and ATRX-2) and 2 healthy controls (Ctr1 and Ctr2).

Combined bisulfite restriction analysis (COBRA)

A total of 100–200 ng of gDNA was chemically or enzymatically converted using EpiTect Fast Bisulfite Conversion Kit (Qiagen) or NEBNext Enzymatic Methyl-seq (NEB), respectively, to detect modified cytosine, in accordance with the manufacturer's instructions. PCR was performed with 15 ng of converted DNA as a template, 0.1 μM primers, 1× TB Green solution, 1× MSP buffer, and 0.6 μl of MSP enzyme in a volume of 20 μL (Takara). Reactions for PCR were subjected to an initial heat denaturation step of 95 °C for 30 s; followed by three cycles of 98 °C for 5 s, 65 °C for 30 s, and 72 °C for 1 min; then three cycles of 98 °C for 5 s, 63 °C for 30 s, and 72 °C for 1 min; and finally 34 cycles of 98 °C for 5 s, 60 °C for 30 s, and 72 °C for 1 min. PCR products purified using AMPure XP beads (Beckman Coulter) were subdivided into two aliquots and treated with or without AccII restriction enzyme (Takara) for 1 h at 37 °C. After electrophoresis in 2.5% agarose gels and ethidium bromide staining, the gel was scanned with a ChemiDoc Touch Imaging System (Bio-Rad) and the band intensities were measured with Image Lab software (Bio-Rad). The percentage of methylation (% methyl) was calculated as the ratio of the cleaved PCR product (methylated) and the total amount of the PCR product (methylated + unmethylated). The following primers were used for PCR: 5′-COBRA_NSD1_F4: GGTT(C/T)GGTGTAGGATGTAGG; COBRA_NSD1_R4: 5′-CTC(A/G)ACACCCAAACAAATAAC. COBRA assay was performed for an independent cohort of 22 Sotos syndrome (SoS10 to SoS31), 9 OGID syndrome (EZH2-1, EZH2-2, EZH2-3, SUZ12-1, SUZ12-2, EED, FOXP1, MTOR and CHD8) and 16 healthy controls (Ctr1 to Ctr16). Four of 22 Sotos syndrome (SoS12, SoS23, SoS26 and SoS27) were removed from the analysis due to the poor PCR amplification using

bisulfite-converted DNA. Sample information is listed in Supplementary Table 2 and 3.

Methylation-specific PCR (MSP)

A total of 100–200 ng of gDNA was bisulfite-converted using the EpiTect Fast Bisulfite Conversion Kit (Qiagen). PCR was performed with 30 ng of converted DNA as a template, 0.3 µM primers, 1 × TB Green solution, 1 × MSP buffer, and 0.48 µl of MSP enzyme in a volume of 20 µL (Takara). Reactions for PCR were subjected to an initial heat denaturation step of 95 °C for 30 s, followed by 37 cycles of 98 °C for 5 s, 55 °C for 30 s, and 72 °C for 1 min. The following methylation-specific primers and unmethylated specific primers were used: ATRX_M_F2: 5′-GGGGCGGCGTAGAATAAAGC; ATRX_M_R2: 5′-TACGAAAAACGAAAACGAC; and ATRX_U_F2: 5′-GGGGTGGTGTAGAATAAAGT; ATRX_U_R2: 5′-TACAAAAAACAAAAACAAC, respectively. MSP was performed for independent cohort of 3 ATR-X syndrome (ATRX-8 to ATRX-10) and 8 healthy controls (Ctr1 to Ctr8). Sample information is listed in Supplementary Table 2.

Reverse-transcription quantitative PCR (RT-qPCR)

Total RNA was extracted from lymphoblastoid cell lines (LCLs) using RNeasy Mini Kit (Qiagen). cDNA was synthesized from 0.5 µg of total RNA and random hexamers using PrimeScript 1st strand cDNA synthesis kit (Takara). To quantify *NSD1* expression, RT-qPCR was performed. Rotor-Gene SYBR Green kit was used for real-time quantification of cDNA, with amplification monitored on the Rotor-Gene cycler system (Qiagen). Target gene expression was compared with *GAPDH* expression as an endogenous control. The following primers were used: NSD1_rt_qF: 5′-GTGACATTAAAGCAGGCACTGA; NSD1_rt_qR: 5′-TTTCTTCCGTGGCAATGGGT; GAPDH_rt_F: 5′-GAAGGTGAAGGTCGGAGTCA; and 5′-GAPDH_rt_R2: GAAGATGGTGATGGGATTTC.

Statistical analysis

Statistical analysis for Fig. 1B was performed using R, version 3.6.2. R function of cor.test with the parameter method = "pearson" was used (https://www.r-project.org/). For Figs. 3E and 5D, Mann–Whitney test was

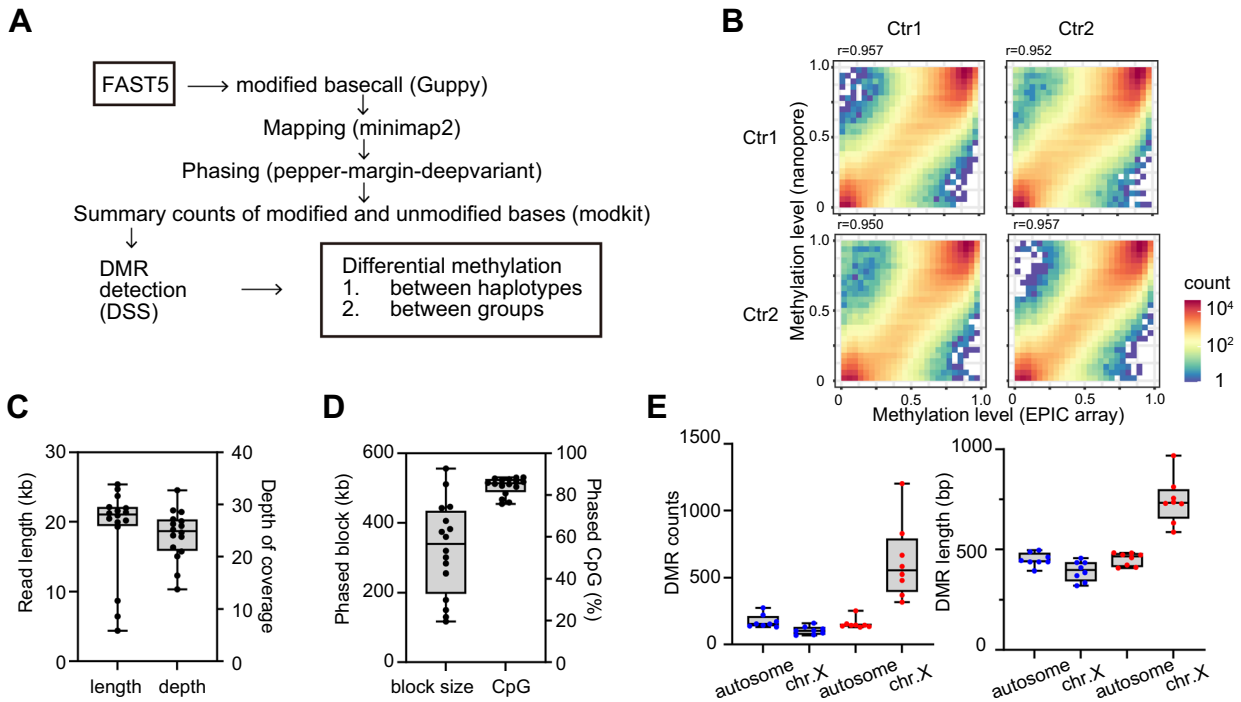

Fig. 1 Overview and performance of the DNA methylation analysis. **A** Flowchart and programs used in this study to detect differentially methylated regions (DMRs) using nanopore long-read sequencing. **B** Pearson correlation coefficient of DNA methylation level between nanopore sequencing and EPIC methylation array. CpGs commonly detected by the two technologies using the same DNA samples (Ctr1 and Ctr2) were compared in all four combinations (comparison of different technologies in the same samples and comparison of different technologies in different samples). **C** The average read length (left) and depth of sequence read coverage (right) distribution of nanopore long-read sequencing datasets from 16 healthy controls. **D** Phased block size in autosomes from 16 healthy controls (left). Percentage of autosomal CpG sites called within phased block (right). **E** The average number of DMRs (left) and the average size of DMRs (right) detected by the DSS program in autosomes and the X chromosome. Males ($n = 8$, shown with blue dots), females ($n = 8$, shown with red dots). Sample information (Ctr1 to 16) is listed in Supplementary Table 1

performed using GraphPad Prism 10 v10.1.2. For Fig. 5C, Kruskal–Wallis test was performed using GraphPad Prism 10 v10.1.2.

Results

Performance of nanopore long-read DNA methylation analysis

We initially evaluated the performance of our nanopore long-read DNA methylation analysis workflow for 5-methylcytosine (5-mC) using the 16 PromethION datasets from healthy controls (Ctr1 to Ctr16) (Supplementary Table 1). The workflow is summarized in Fig. 1A. 5-mC was called from raw nanopore data (FAST5 files) using Guppy with a modified basecalling model that was trained to distinguish 5-mC from unmethylated cytosine [21]. Nanopore reads for which 5-mC information was available were mapped to the reference human genome GRCh38 and phased using minimap2 [16] and pepper-margin-deepvariant [17], respectively. 5-mC information was extracted and manipulated from the resultant phased bam files (aligned bam files with both 5-mC and haplotype information) using modkit. We used the DSS program [19] to perform differential methylation analysis between two haplotypes (allele-specific change), or between two groups (i.e., cases and controls, to detect possible disease-associated change).

We first compared the methylation level at each CpG site between nanopore sequencing and well-established methylation array data from the same DNA samples (Ctr1 and Ctr2). The nanopore sequencing showed a high correlation with the methylation array with a Pearson correlation coefficient of 0.95 (comparison of common 823,621 or 825,812 autosomal CpGs between the two technologies for Ctr1 and Ctr2, respectively) (Fig. 1B). In addition, the nanopore sequencing fully covered the epigenome; that is, it called 98.0% (26,354,826/26,900,281) or 98.2% (26,428,342/26,900,281) of autosomal CpGs compared to a rate of 3% (829,207/26,900,281 or 829,207/26,900,281) in the methylation arrays in Ctr1 and Ctr2, respectively, as expected (Supplementary Fig. 1A, B). We also note that nanopore sequencing does not suffer from high/low GC-content bias, which is often problematic in short-read NGS data, as indicated by a tight CpG coverage distribution at around 20×sequence reads (Supplementary Fig. 1C, D). We extended our analysis to 16 control samples (Ctr1 to Ctr16) using the workflow shown in Fig. 1A and assessed our datasets. The median of the average read length and median of the average depth of coverage in our 16 datasets were 21,065 bp (range 4385–25,388) and 24.89×(range 13.78–32.72), respectively (Fig. 1C). This long-range DNA sequence information with 5-mC signals allowed us to perform haplotype-aware methylation analysis. To assess the effectiveness of genome-wide haplotype-aware differential methylation analysis, we calculated the haplotype phased block size and fraction of CpGs phased by these haplotypes. The median of the average phased block size was 339,782 bp (range 117,284–556,202, $n=16$), and these phase blocks covered 84.2% (range: 75.8% to 88.5%, $n=16$) of autosomal CpGs from the 26,900,281 CpGs in GRCh38 used in this study (Fig. 1D). As expected, a significant number of DMRs were called in females for the X chromosome (median of average DMRs: 103 and 553 per sample in 8 males and 8 females, respectively), indicating random X chromosome inactivation by differential methylation analysis between two haplotypes (HP1: Haplotype 1 and HP2: Haplotype 2) (Fig. 1E) [22]. Such X chromosome-associated DMRs in females were also larger (median of average DMR length: 397 bp and 733 bp in 8 males and 8 females, respectively), which may reflect the sum of both hypermethylation of promoters and hypomethylation of gene bodies on the inactive X chromosome (both can be detected as DMRs) [23] (Fig. 1E). These observations suggest that our nanopore sequencing workflow provides reliable 5-mC profiles throughout the entire genome.

Advantage for detecting DMRs in nanopore sequencing

In addition to random X chromosome inactivation, genome imprinting is a different type of regulatory mechanism of monoallelic gene expression based on the parent of origin [24]. Many parent-of-origin-derived imprinting DMRs (iDMRs) have been reported [20, 25–27], and dysregulation of these iDMRs has been implicated in human diseases [28]. The iDMRs were generally recognized as partially methylated regions using previous technologies with haplotype-unphased datasets. In contrast, nanopore sequencing with a DNA methylation profile provides the opportunity to directly compare the methylation level based on two haplotypes. Our analysis detected 3976 DMRs on average between two haplotypes (HP1 and HP2) at the genome-wide level (range 2939–6364, $n=16$). Among them, 88% of well-characterized iDMRs were correctly detected in one representative dataset (Ctr11) (82 of 93 iDMRs from multiple studies, from the work of Akbari et al.) [29], whereas relatively minor iDMRs reported in only a single study were rarely detected (32%, 32 out of 99 iDMRs from a single study) [29], which may have been related to the tissue types examined and/or inter-individual variations (Fig. 2A–C, Supplementary Table 4) [30]. We note that *H19* DMR, which was shown to be an important imprinting domain for Beckwith–Wiedemann syndrome and Silver–Russell syndrome [31], within a 99,421 bp segmental duplication with 0.9988 sequence identity, was called (Fig. 2A) and a long-range DNA methylation profile between two haplotypes was clearly demonstrated (Fig. 2B). This

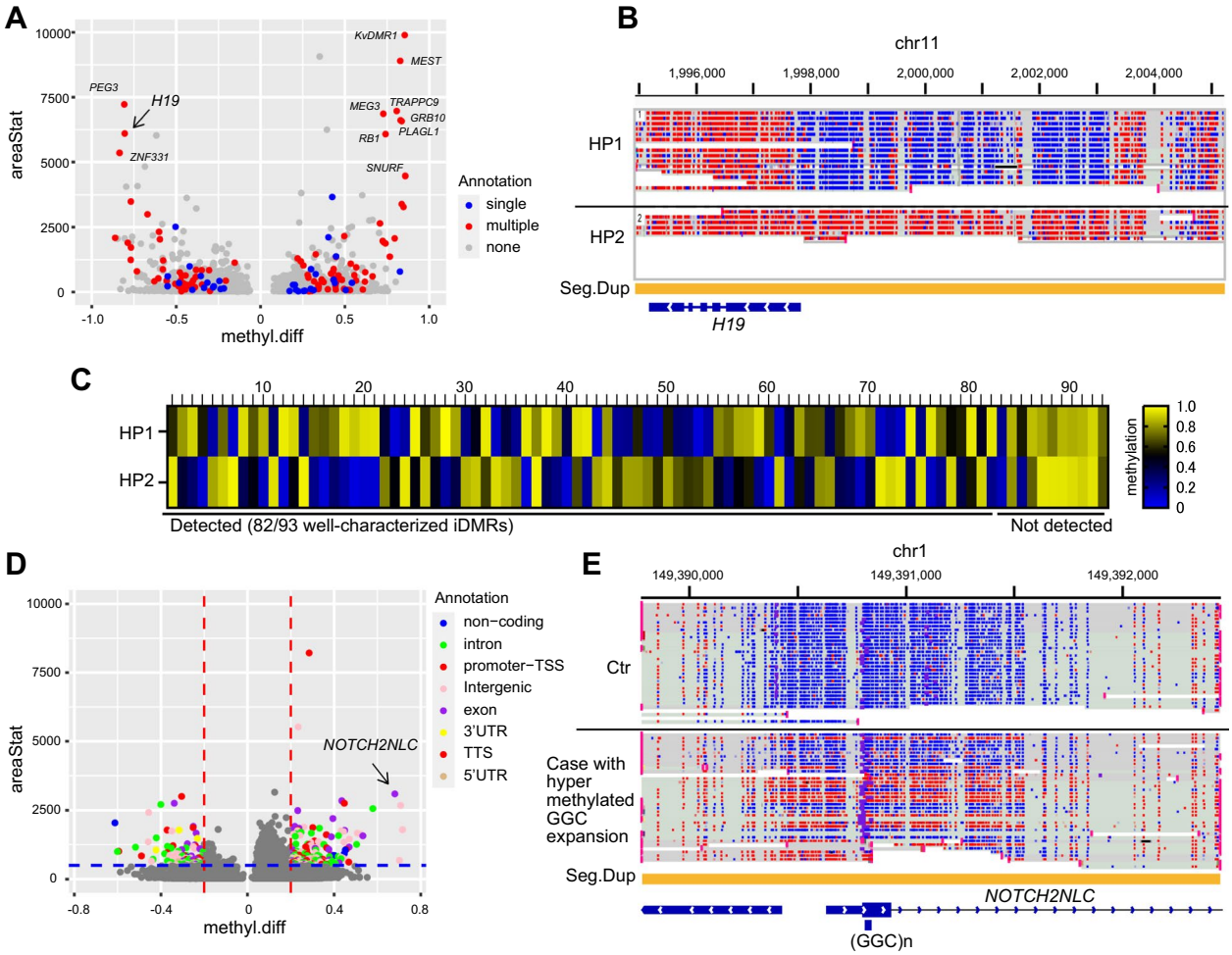

Fig. 2 DMR detection in difficult-to-sequence regions. **A** Genome-wide screening of DMRs between two haplotypes using DSS in Ctr11. DMR calls that overlapped with well-characterized imprinting DMRs (iDMRs) from multiple studies and iDMRs from single studies are annotated in red and blue, respectively. The difference in methylation levels between haplotypes (X-axis) and areaStat parameter (Y-axis) from DSS is plotted. Y-axis is limited to (0,10,000) and two datapoints with high areaStat values have been removed to aid visualization. Arrow indicates *H19* iDMR shown in Fig. 2B. **B** IGV visualization of nanopore reads for *H19* iDMR in Ctr11. Haplotype-aware long-range methylation analysis revealed unmethylated CpG (blue) and methylated CpG (red) in haplotypes 1 (HP1) and 2 (HP2), respectively. The orange rectangle represents segmental duplication (Seg. Dup). **C** Heatmap visualization of methylation levels of two haplotypes (HP1 and HP2) at 96 well-characterized iDMRs in Ctr11. **D** Genome-wide screening of DMRs in a case of extreme *NOTCH2NLC* GGC repeat expansion (NOTCH2NLC_F3_father) and 8 healthy controls (Ctr1 to Ctr8). DMRs detected by DSS are plotted as shown in Fig. 2A. Y-axis is limited to (0,10,000) and two datapoints with high areaStat values have been removed to aid visualization. Arrow indicates aberrant DNA hypermethylation with *NOTCH2NLC* GGC repeat expansion. Regions with an absolute value of areaStat greater than 500 (|areaStat|> 500, blue dashed line) and an absolute value of methylation difference greater than 0.2 (|methylation difference|> 0.2, red dashed lines) are considered as significant DMRs. DMRs were annotated in the context of genome annotation by the Homer program (see methods). TSS: transcription start site, UTR: untranslated region, TTS: transcription termination site. **E** IGV visualization of nanopore reads of heterozygous *NOTCH2NLC* GGC repeat expansion with hypermethylated CpG (red), whereas non-expanded alleles are hypomethylated (blue) in a case of extreme GGC expansion of *NOTCH2NLC* and in Ctr. The orange rectangle represents segmental duplication (Seg.Dup). Ctr: healthy control. Sample information is listed in Supplementary Table 1

suggested the opportunity to expand our analysis to genomic regions at which sequencing is generally problematic [32]. To test this possibility further, we studied GGC repeat expansion of *NOTCH2NLC*, a gene causative of neuronal intranuclear inclusion disease [33, 34], with DNA methylation alteration depending on the size of the GGC repeats. *NOTCH2NLC* is located at a segmental duplication and GGC repeat expansion exceeding 300 repeats at this locus tends to be CpG hypermethylated, as shown by targeted long-read sequencing [13]. We tested whether unbiased genome-wide screening could correctly extract this as a significant DMR. Indeed,

genome-wide comparison between a case of extreme GGC expansion of *NOTCH2NLC* (NOTCH2NLC_F3_father in Supplementary Table 1) [13] and healthy controls (Ctr1 to Ctr8) successfully identified *NOTCH2NLC* GGC repeats as the eighth ranked candidate from 12,979 possible DMRs using DSS (Fig. 2D, E, Supplementary Table 5). Consideration of the genomic annotation for these DMR candidates may help to rank disease-relevant DMRs from a large number of DMR candidates, as indicated in Fig. 2D and Supplementary Table 5. This highlights *NOTCH2NLC* as an exonic DMR, which is known to regulate aberrant *NOTCH2NLC* expression leading to toxic polyglycine and/or RNA products [35].

In summary, long-read sequencing can provide the opportunity for the haplotype-aware and repeat-resolved methylation analysis with nearly complete sets of CpGs.

Application of nanopore sequencing for DNA methylation signatures in rare disease

Recent studies have demonstrated the diagnostic utility of DNA methylation signatures in some NDDs, particularly for syndromes with overlapping clinical features [3, 4]. Given the performance and advantages mentioned above (Figs. 1, 2), long-read sequencing with DNA methylation profiling may offer benefits for research and diagnosis using DNA methylation signatures. To test this possibility, we selected Sotos syndrome (OMIM# 117,550) and alpha thalassemia/mental retardation X-linked syndrome (ATR-X syndrome, OMIM# 301040). Aberrations of genes encoding the histone methyltransferase NSD1 and chromatin remodeler ATRX cause Sotos syndrome and ATR-X syndrome, respectively, and the presence of syndrome-specific methylation signatures in these conditions is well accepted [36, 37]. Analysis overview for DNA methylation signatures is summarized in Supplementary Fig. 2. Two previous EPIC array-based studies reported almost distinct sets of DNAm signatures, 37 and 112 loci for Sotos syndrome and 47 and 101 loci for ATR-X syndromes [6, 7]. Initially, we generated LR-GS data for six samples along with EPIC methylation array experiments using the same DNA (SoS1, SoS2, ATRX-1, ATRX-2, Ctr1 and Ctr2) to compare two technologies (EPIC methylation array and LR-GS) and evaluated these reported DNAm signatures. Methylation level of each CpG site of reported DNAm signatures for Sotos and ATR-X syndromes were extracted from both EPIC methylation array and LR-GS data (37 and 47 CpGs for Sotos and ATR-X in Supplementary Fig. 3A and 3C, 112 and 101 CpGs for Sotos and ATR-X in Supplementary Fig. 3B and 3D) [6, 7]. As expected, unique DNA methylation patterns were observed in the heatmap visualization (Supplementary Fig. 3). Similar syndrome-specific methylation profiles of Sotos and ATR-X syndromes were confirmed by multi-dimensional scaling (MDS), as indicated by the discrete clustering of three groups: Sotos syndrome, ATR-X syndrome, and healthy controls (Supplementary Fig. 3), suggesting the compatibility of long-read based methylation assay with EPIC array-based DNA methylation signature. Despite this compatibility, we noticed that there is a relatively large variation of methylation level at respective CpG sites due to the read count-based scoring of methylation level using the low and uneven read coverage datasets (Supplementary Fig. 3C and 3D). We reasoned that the regional analysis after merging neighboring CpGs into a single region could be preferable for LR-DNAm signatures to compensate this issue. Hence, we used the DSS program to find differential methylation regions between two groups (i.e., Sotos, ATR-X syndromes and controls) as described above. The DNA methylation levels of three molecularly diagnosed cases of respective Sotos syndrome (SoS1, SoS2, and SoS3) and ATR-X syndrome (ATRX-1, ATRX-2, and ATRX-3) were compared with those of eight controls (Ctr1 to Ctr8) using a LR-GS data at autosomal CpGs at the genome-wide scale. Overall, 24,522 and 9,449 DMR candidates were called in Sotos syndrome and ATR-X syndrome, compared with healthy controls (Fig. 3A, B), respectively. To check the sensitivity of the DMR detection by DSS, we checked the extent of the overlap between our 24,522 DMR candidates and 7,033 Sotos syndrome-associated differentially methylated loci from EpigenCentral [5]. Among 6,531 of 7,033 probes, which were successfully converted from hg19 to GRCh38 human reference genome, 85.3% (5571/6531 probes) is covered by 24,522 DMRs detected by our LR-GS data analysis, suggesting the reasonable coverage of Sotos syndrome-associated differentially methylated loci. However, the number of candidates called by DSS (i.e., 24,522 and 9,449 DMR candidates in Sotos syndrome and ATR-X syndrome, respectively) with the default parameters was too high to evaluate, and thus we narrowed down the candidates by ad hoc classification. Because DMRs with a larger areaStat value (a parameter from DSS based on the statistics of Wald test, height and width of DMRs) [19] and a higher degree of methylation difference are more likely to be reliable, we selected regions in which the absolute value of areaStat is greater than 500 ($|\text{areaStat}| > 500$) and the absolute value of methylation difference is greater than 0.2 ($|\text{methylation difference}| > 0.2$) as significant DMRs. Upon applying this classification to analyze the control dataset in Fig. 2A and D, we were able to narrow down the candidates from 3,144 to 255 (data for Fig. 2A) and from 12,979 to 284 (data for Fig. 2D), constituting 91.9% and 97.8% reductions, respectively. This filtering still retained 66% (54 of 82) of the well-characterized iDMRs and *NOTCH2NLC* repeat expansion-associated

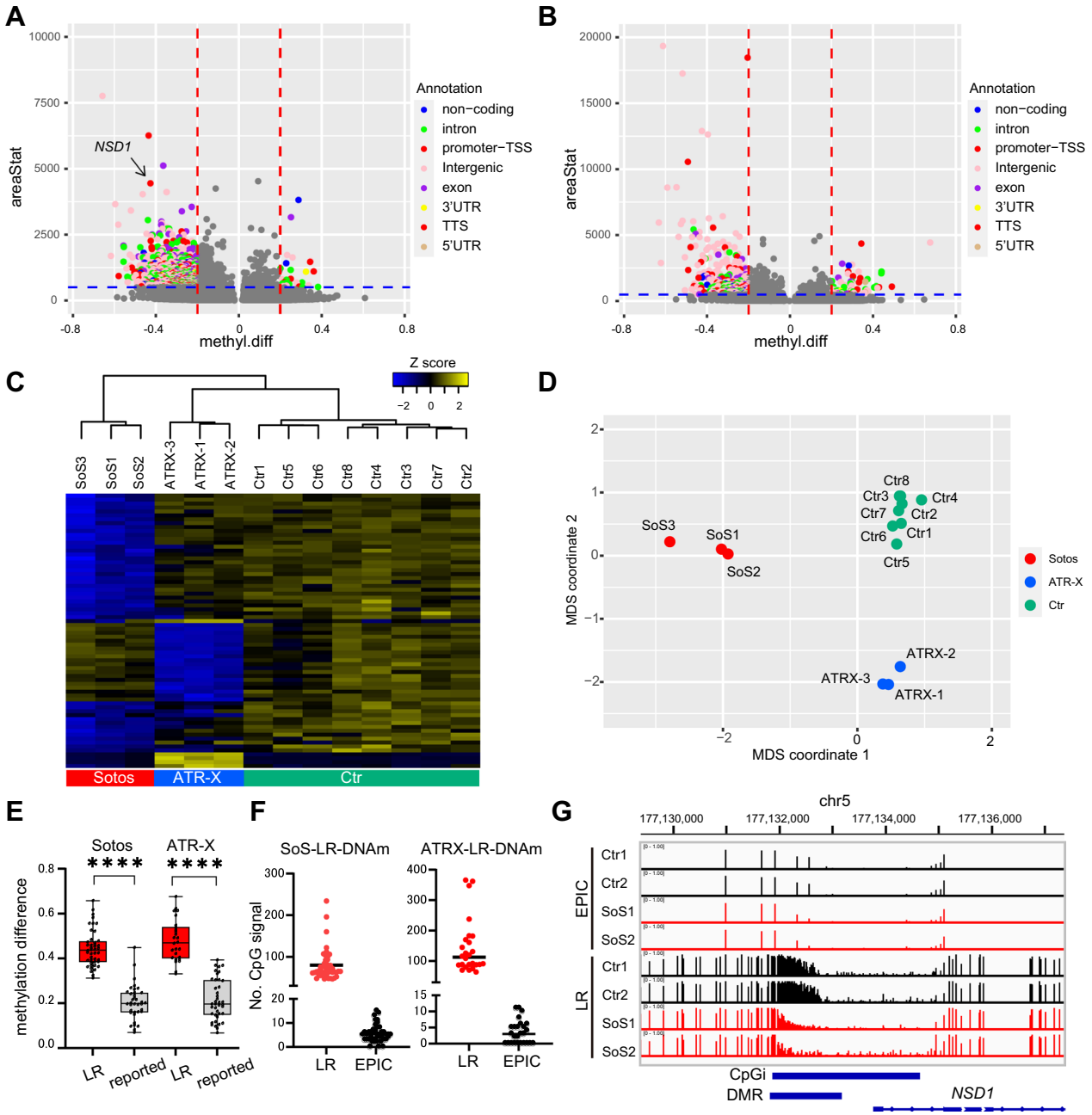

Fig. 3 DNAm signature identified by nanopore long-read sequencing. **A, B** Genome-wide screening of DMRs in Sotos syndrome (A) or ATR-X syndrome (B). Y-axis is limited to (0,10,000) and one and two datapoints with high areaStat values have been removed to aid visualization in (A) and (B), respectively. DMRs are plotted as in Fig. 2D. Arrow indicates *NSD1* CpGi shown in Fig. 3G in (A). **C** Heatmap visualization of sets of differential methylation marks for Sotos syndrome and ATR-X syndrome (Sotos and ATR-X LR-DNAm signatures). The methylation level is colored from yellow (indicating high DNA methylation) to blue (indicating low DNA methylation). Sotos syndrome ($n=3$, SoS1 to SoS 3), ATR-X syndrome ($n=3$, ATRX-1 to ATRX-3), and healthy controls ($n=8$, Ctr1 to Ctr8) are labeled red, blue, and green, respectively, in the bottom. **D** Multi-dimensional scaling (MDS) plot showing the similarity of DNA methylation pattern in Sotos syndrome and ATR-X syndrome. Each point represents one subject colored red, blue, and green for Sotos ($n=3$), ATR-X ($n=3$), and healthy controls ($n=8$), respectively. **E** The median methylation difference at LR-DNAm signatures and reported DNAm signatures. Sotos LR-DNAm ($n=44$ DMRs); reported DNAm signatures ($n=36$ DMRs); ATR-X LR-DNAm signatures ($n=26$ DMRs); 40 reported DNAm signatures ($n=40$ DMRs). ****: significant differences with p-value < 0.0001 by Mann–Whitney test. **F** The number of CpGs available for analysis within regions of Sotos and ATR-X LR-DNAm signatures by nanopore sequencing and EPIC methylation array in the same DNA samples (SoS1 and ATRX-1). Each point represents a region of 44 Sotos and 26 ATR-X LR-DNAm signatures. Black bars represent mean values of the number of CpGs. **G** DNA methylation profiles at *NSD1* CpGi DMR by two different technologies: nanopore sequencing and EPIC methylation array. Data were obtained from the same DNA samples (Ctr1, Ctr2, SoS1, and SoS2). CpGi: CpG island annotation from UCSC (http://genome.ucsc.edu/); DMR: DMR detected by DSS. *NSD1*: NM_022455.5. Sample information is listed in Supplementary Table 1

DNA hypermethylation for Fig. 2A and D, respectively. Although not perfect, we consider these DMRs to be biologically significant DMR candidates. On the basis of this classification, 1,111 and 427 DMRs were classified as significant DMRs for Sotos and ATR-X syndromes, respectively (Fig. 3A, B). Interestingly, only 37.8% (14 of 37) [7] or 55.3% (62 of 112) [6] and 25.5% (12 of 47) [7] or 19.8% (20 of 101) [6] of the reported Sotos- and ATR-X-associated DNAm signatures are overlapped with significant LR-based DMRs, respectively (Supplementary Fig. 4). Hence, we wondered whether alternative DNAm signatures could be identified from other significant DMRs detected by nanopore sequencing. Considering sequencing depth, size of DMRs, and inter-individual variations, 44 of 1,111 significant DMRs and 26 of 427 significant DMRs were selected by manual curation for Sotos syndrome and ATR-X syndrome, respectively (we refer to these as Sotos LR-DNAm and ATR-X LR-DNAm signatures in this manuscript) (Fig. 3C, Supplementary Tables 6, 7). Similarity of the methylation profiles of Sotos syndrome and ATR-X syndrome using 44 Sotos- and 26 ATR-X LR-DNAm signatures was confirmed by multi-dimensional scaling (MDS), as indicated by the discrete clustering of three groups: Sotos syndrome, ATR-X syndrome, and healthy controls (Fig. 3D). Furthermore, these LR-DNAm signatures were also confirmed by an orthogonal experimental approach (using EPIC methylation array data) (Supplementary Fig. 5).

Next, we characterized LR-DNAm signature (44 Sotos LR-DNAm and 26 ATR-X LR-DNAm signatures) in comparison with reported microarray-based DNAm signatures. Among the 37 and 47 probes of reported microarray-based DNAm signatures for Sotos and ATR-X syndromes [7], respectively, 36 and 40 regions containing these probes are called as DMRs in our LR-based DMR detection analysis using DSS. The methylation difference in LR-DNAm signatures (mean 0.44 methylation difference with range of 0.31–0.66 and 0.35 with range of 0.27–0.50 for Sotos LR-DNAm and ATR-X LR-DNAm signatures, respectively) was greater than that of reported array-based DNAm signatures [mean 0.20 with range of 0.07–0.45 (36 of 37 DMRs detected by DSS) and mean 0.19 with range of 0.07–0.42 (40 of 47 DMRs detected by DSS) from reported DNAm signatures for Sotos and ATR-X, respectively] (Fig. 3E).

As mentioned above, nanopore sequencing fully covered the epigenome, indicating that a higher number of CpGs could enhance the detection of LR-DNAm signatures. In fact, 44 and 26 regions of differential methylation marks for Sotos and ATR-X syndromes (LR-DNAm), respectively, were covered by a higher number of CpGs in nanopore sequencing compared with that with a methylation array (median of 65.0 CpGs with a range of 47–234 in nanopore sequencing versus 5.0 CpGs with a range of 0–15 in the methylation array for Sotos LR-DNAm signature region; and median of 112.5 CpGs with a range of 64–366 in nanopore sequencing versus 3.0 CpGs with a range of 0–11 in the methylation array for the ATR-X LR-DNAm signature region) (Fig. 3F). One interesting example is hypomethylation of a part, but not all of the *NSD1* CpGi in Sotos syndrome (chr5: 177131773–177133000) (Fig. 3G). A total of 122 CpGs were called for nanopore sequencing, whereas 5 CpGs present on the methylation array. This partial *NSD1* CpGi DMR was confirmed by methylation array using the same DNA samples (Ctr1, Ctr2, SoS1, SoS2), showing differential methylation β values (% of methylation from the probe intensity signal) [4] at three of five CpG probes (cg19731612, cg18121224, and cg08369368 with p-values of 0.0007, 0.0016, and 0.0150 by Welch's t-test, respectively) (Fig. 3G), two of which were also reported in EpigenCentral (cg19731612 and cg18121224) [5, 37].

Validation of long-read DNA methylation signatures (LR-DNAm signatures)

To effectively screen LR-DNAm signatures from the perspectives of time, cost, computer resources, and data storage, we applied nanopore adaptive targeted long-read sequencing (adaptive T-LRS) for 70 combined genomic regions of Sotos and ATR-X LR-DNAm signatures (Supplementary Tables 6, 7) using nanopore GridION sequencer (Fig. 4A). Using adaptive T-LRS or LR-GS, 24 independent samples (6 Sotos syndrome, 4 ATR-X syndrome, 8 healthy controls, and 6 other NDDs) were studied, and Sotos and ATR-X LR-DNAm signatures were validated (Supplementary Table 1). Six, four, and eight individuals were correctly assigned to the Sotos syndrome, ATR-X syndrome, and healthy groups, respectively, in MDS analysis. Moreover, six individuals with other NDDs, whose molecular diagnosis was negative for *NSD1* or *ATRX* pathogenic variants, were not grouped into Sotos syndrome or ATR-X syndrome (Fig. 4B, C, Supplementary Table 1), suggesting unique LR-DNAm signatures for Sotos syndrome and ATR-X syndrome. Next, we built a support vector machine (SVM) model to estimate the probability that individuals fall into the disease class (Sotos syndrome or ATR-X syndrome). We created the SVM model using a discovery cohort ($n=14$ datasets) (Supplementary Fig. 6) and made predictions about disease category on a validation cohort ($n=24$ datasets). This model classified Sotos syndrome, ATR-X syndrome, and controls (healthy or other NDD groups) correctly with a methylation variant pathogenicity (MVP) prediction score of 0.5 or higher (Fig. 4D) [4].

Three samples with ATR-X (ATRX-8, 9, and 10 in Supplementary Table 2) could not be tested for the global

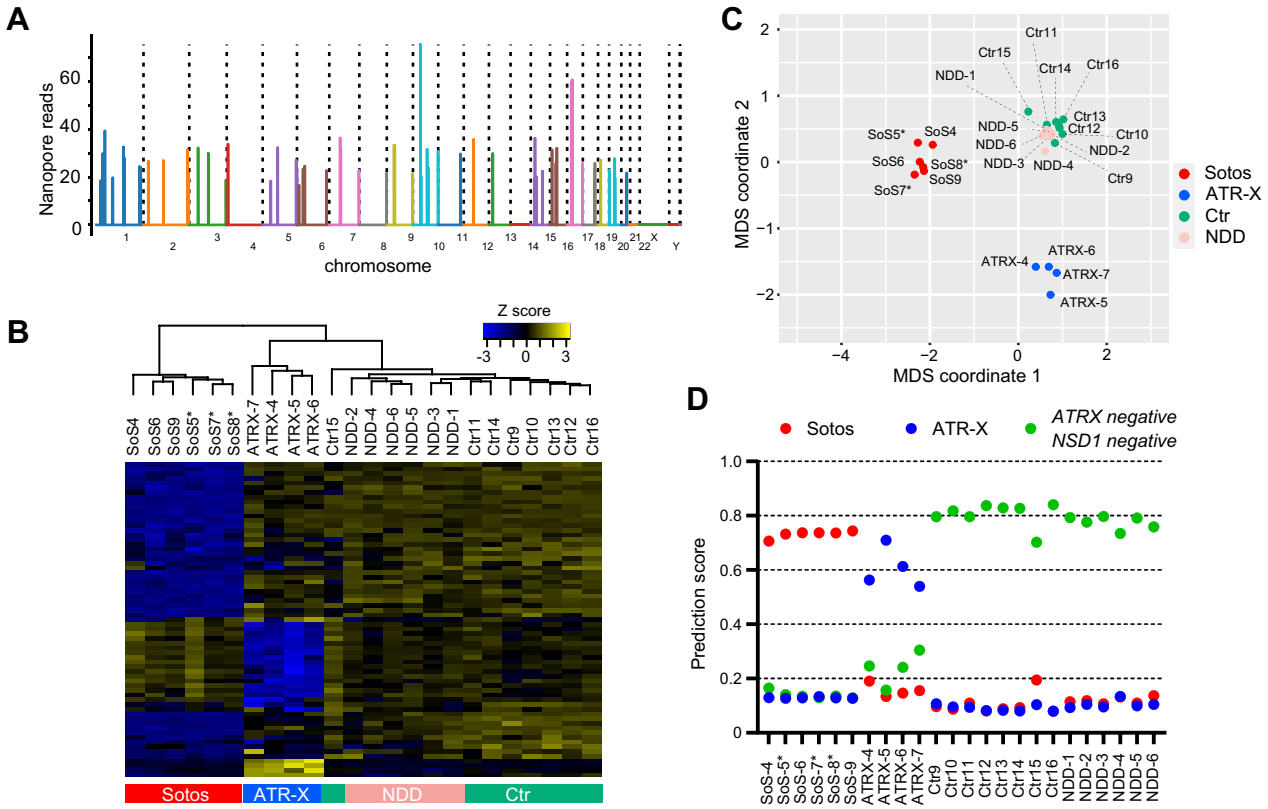

Fig. 4 LR-DNAm signatures in the validation cohort. **A** Coverage plot of adaptive T-LRS targeting genomic regions of 44 Sotos and 26 ATR-X LR-DNAm signatures (70 regions in total). Data from one representative sample (SoS6) are shown. X-axis: genomic position. Y-axis: number of on-target nanopore reads. **B** Heatmap visualization of LR-DNAm signatures in validation cohort. Sotos ($n=6$, SoS4 to SoS 9), ATR-X ($n=4$, ATRX-4 to ATRX-7), healthy controls ($n=8$, Ctr9 to Ctr16), and other NDDs ($n=6$, NDD-1 to NDD-6) are labeled red, blue, green, and pink, respectively, in the bottom. Three cases of Sotos syndrome harboring a 5q35 microdeletion (involving *NSD1*) are annotated by asterisk (SoS5, SoS7 and SoS8). **C** MDS plot showing the similar DNA methylation patterns in the validation cohort. Each point represents one subject colored red, blue, green, and pink for *NSD1* ($n=6$), *ATRX* ($n=4$), healthy controls ($n=8$), and other NDDs ($n=6$), respectively. Three cases of Sotos syndrome harboring a 5q35 microdeletion are annotated by asterisk (SoS5, SoS7 and SoS8). **D** Prediction based on the LR-DNAm signatures using SVM in the validation cohort. The probability scores that each individual belongs to the Sotos syndrome, ATR-X syndrome, and *NSD1*- and *ATRX*-negative groups are shown in red, blue, and green dots, respectively. Sample information is listed in Supplementary Table 1

multi-locus ATR-X LR-DNAm signature by nanopore sequencing due to the limited amount of DNA. For these three samples, we selected one representative DMR (ATRX LR-DNAm-031) and confirmed differential methylation by a single-locus methyl-specific PCR (MSP) assay (Supplementary Fig. 7).

Diagnostic utility of DNA hypomethylation of partial *NSD1* CpGi in both cases with *NSD1* point mutations and those with a 5q35 deletion

As mentioned in Fig. 3G, we unexpectedly identified hypomethylation of partial *NSD1* CpGi in Sotos syndrome. Considering that Sotos syndrome is caused by haploinsufficiency of *NSD1*, cells with hypomethylated alleles may have some selective growth advantage by compensating for the decreased expression of NSD1 protein. Actually, this is not the case at least in lymphoblastoid cell lines (LCLs), as indicated by normal or decreased expression of *NSD1* in a case having a hypomethylated allele (SoS1) using real-time quantitative RT-PCR assays (Supplementary Fig. 8), which is consistent with a previous report [38]. Determination of the 5-mC status and *NSD1* expression in the organs mainly affected by Sotos syndrome, such as skeletal, cardiovascular, brain, and urinary systems, might be interesting. Nevertheless, testing this DNA methylation mark might be useful as a diagnostic marker. Hence, we screened the differential methylation at a part of *NSD1* CpGi using combined bisulfite restriction analysis (COBRA) [39]. We initially validated this in a patient–parents trio. In SoS1 with a de novo *NSD1* pathogenic missense variant (NM_022455.5:c.5885T>C:p.Ile1962Thr), hypomethylation at *NSD1* CpGi was only present in the patient, but not in his parents (Fig. 5A). Sotos syndrome is caused by

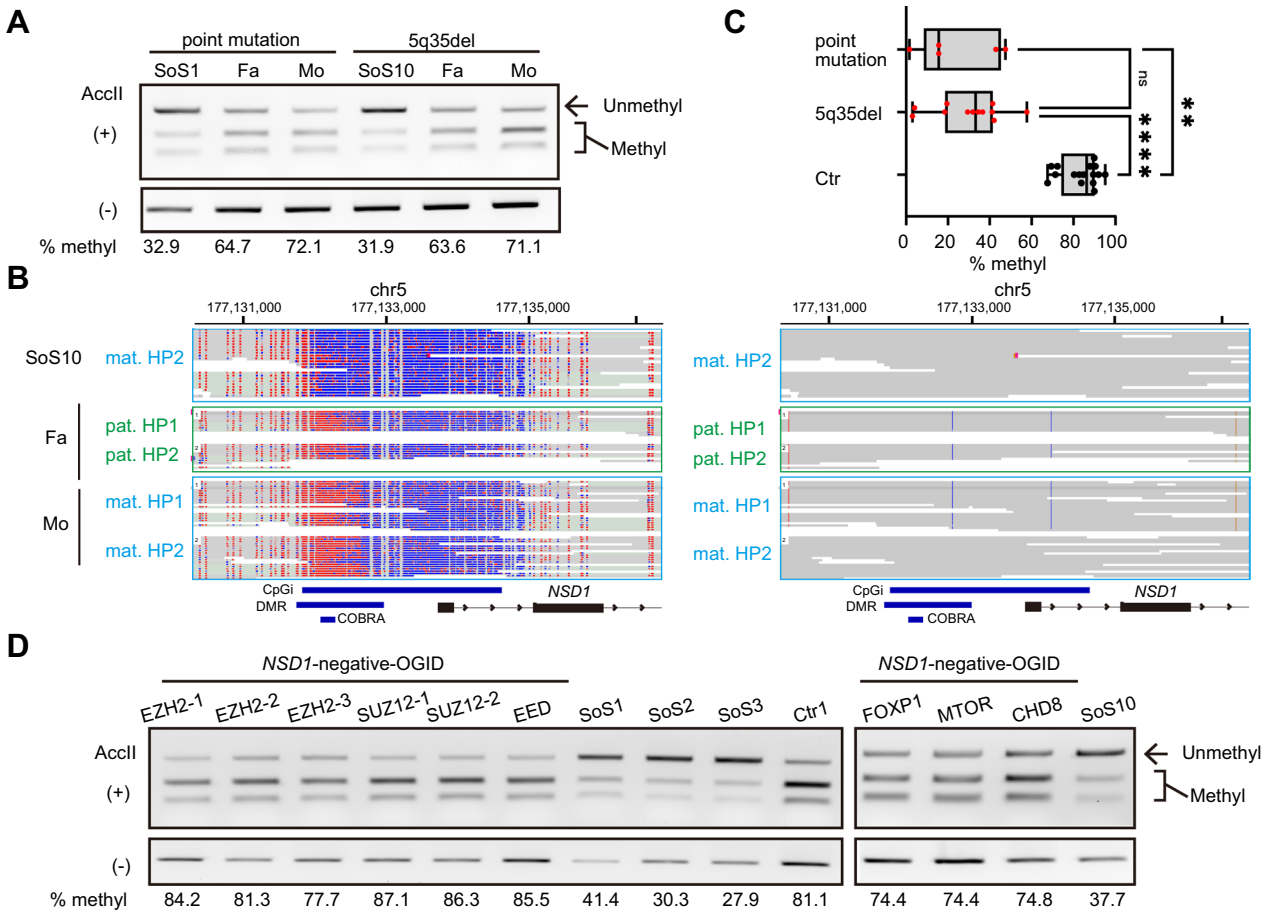

Fig. 5 *NSD1* CpGi DMR in cases of *NSD1* point mutations and 5q35 submicroscopic deletions. **A** Confirmation of differential methylation using conventional PCR-based COBRA assay in two families: SoS1 with p.Ile1962Thr and SoS10 with a 5q35 deletion, along with the parents. PCR products were digested with (+) or without (–) AccII. The methylation level (% methyl) was calculated using the ratio of Methylated fraction / (Methylated + Unmethylated fraction) from the gel image. **B** IGV visualization of PacBio HiFi reads at *NSD1* CpGi DMR in the SoS10 family. Haplotype phasing revealing the origin of the hypomethylated chromosome (i.e., maternal HP2). HiFi reads were grouped by haplotype information (right). Multiple SNVs (colored bars) on HiFi reads can be used to construct regional haplotypes. Green and blue boxes represent paternal and maternal chromosomes, respectively. Fa: father; Mo: mother; CpGi: CpG island annotation from UCSC (http://genome.ucsc.edu/); DMR: *NSD1* CpGi DMR detected by DSS; COBRA: amplicon by the COBRA primers. *NSD1*: NM_022455.5. **C** Comparison of the methylation level based on the COBRA assay among three groups: *NSD1* point mutations or a single-exon deletion ($n = 5$), 5q35 submicroscopic deletions ($n = 12$), and healthy controls ($n = 16$). ns: not significant; ****: significant differences with p-value < 0.0001 by Kruskal–Wallis test; **: significant differences with p-value of 0.0021 by Kruskal–Wallis test. **D** Comparison of the methylation level based on the COBRA assay between Sotos syndrome and other overgrowth intellectual disability syndromes (OGIDs). PCR products were digested with (+) or without (–) AccII. The methylation level (% methyl) was calculated by the ratio of Methylated fraction/(Methylated + Unmethylated fraction) from the gel image. Sample information is listed in Supplementary Tables 2 and 3

both *NSD1* point mutations and 5q35 submicroscopic deletions involving the entirety of *NSD1* [40]. We found that cases of a 5q35 submicroscopic deletion (SoS5, SoS7, and SoS8) had the same Sotos LR-DNAm signature as cases of *NSD1* point mutations in validation cohorts (Fig. 4B, C, D). Consistent with this, patient- and locus-specific methylation difference at *NSD1* CpGi was confirmed in another family with a 5q35 submicroscopic deletion (SoS10) using COBRA assay (Fig. 5A). This indicates that the DNA methylation status of the intact allele, not the mutated (deleted) allele, could be used as a diagnostic marker. To confirm this, we performed haplotype phasing analysis using PacBio HiFi sequencing, rather than relatively noisy nanopore long-read sequencing, in the SoS10 family. Haplotype phasing with DNA methylation information clearly revealed that the 5q35 microdeletion was of paternal origin (Supplementary Fig. 9), and the hypermethylated allele in the patient's mother turned out to be hypomethylated after being inherited by an affected child (mat.HP2 in Fig. 5B). These observations

led us to investigate whether a single-locus methylation mark at *NSD1* CpGi could be used for both cases of *NSD1* point mutations and cases of a 5q35 submicroscopic deletion. To test this possibility, we recruited 5, 1, and 15 patients molecularly diagnosed with Sotos syndrome with *NSD1* point mutations, a single-exon deletion, and 5q35 submicroscopic deletions, respectively ($n=21$ in total) (Supplementary Table 2). PCR amplification was not efficient in four DNA samples (SoS12, 23, 26, and 27), and these samples were removed from quantitative methylation frequency analysis using COBRA (Supplementary Table 2). This analysis showed a statistically significant difference in methylation between *NSD*1 point mutations/a single-exon deletion and healthy controls (mean 24.7% methylation with a range of 1.6–47.5% vs. mean 83.5% with a range of 67.6–95.1%), as well as between 5q35 submicroscopic deletions (mean 30.0% methylation with a range of 3.0–57.7%) and healthy controls (Fig. 5C), supporting the diagnostic utility of the hypomethylation of *NSD1* CpGi.

Methylation status of *NSD1* CpGi in other overgrowth intellectual disability syndromes

Sotos syndrome is characterized by a distinctive facial appearance, intellectual disability, and overgrowth. Such clinical features are occasionally obscured depending on the patient's age and several other clinical conditions should be considered for differential diagnosis if clinically overlapping phenotypes are seen (overgrowth intellectual disability syndromes: OGIDs) [41]. We wondered whether these OGIDs have the same DNA methylation alteration at *NSD1* CpGi. We tested this possibility in individuals consulting at other hospitals with a clinical suspicion of Sotos syndrome, Weaver syndrome, or overgrowth with unknown etiology, but negative for *NSD1* pathogenic variants (*NSD1*-negative OGIDs). Nine individuals with *NSD1*-negative OGID with pathogenic variants of *EZH2* ($n=3$, OMIM# 277590), *SUZ12* ($n=2$, OMIM# 618786), *EED* ($n=1$, OMIM# 617561), *FOXP1* ($n=1$, OMIM# 613670), *MTOR* ($n=1$, OMIM# 616638), and *CHD8* ($n=1$, OMIM# 615032) were tested (Supplementary Table 3). Despite features that overlap with Sotos syndrome, along with age and sex, these samples were negative for hypomethylation at *NSD1* CpGi by COBRA assay [median 34.0% ($n=4$, Sotos syndrome) versus 81.3% ($n=9$, *NSD1*-negative OGIDs), with a significance difference (p-value 0.0028 by Mann–Whitney test); Fig. 5D], suggesting the utility of a single-locus DNA methylation test to differentiate Sotos syndrome from other OGIDs.

Discussion

Recent improvements of the read quality and yield of LR-GS have enabled the comprehensive discovery of variants, including SNVs, SVs, complex chromosomal abnormalities, and repeat expansion mutations [10]. In this study, we examined whether DNA methylation analysis using long-read sequencing can be implemented as part of a comprehensive genetic test for rare diseases. Several benefits and drawbacks of this approach were identified in this study, as described below.

In terms of the advantages, LR-GS can simultaneously detect DNA sequences and DNA methylation without any extra epigenetic analyses from nascent DNA without chemical or enzymatic base conversion. As a result, the analytical process is simple and provides highly comparable genetic and epigenetic information from a single dataset (Figs. 1A, 5B), whereas microarray-based and short-read-based DNA methylation analysis require additional (separated) epigenetic experiments, which are time-consuming and labor-intensive [42]. The analytical results also have a high resolution (Supplementary Fig. 1A, B) with uniform coverage (Supplementary Fig. 1C, D). This approach also enables haplotype-aware (80–87% of CpG, Fig. 1D and Fig. 2A–C) and repeat-resolved (including segmental duplications, Fig. 2B, E) analyses with nearly complete sets of CpGs (Supplementary Fig. 1A, B). All of these advantages enable us to successfully detect LR-DNAm signatures, as alternative sets of markers to those obtained by methylation arrays (Figs. 3 and 4). In fact, these alternative sets of markers were covered by a higher number of methylation calls (Fig. 3F, G) and some of these overlapped with segmental duplications [25% of *NSD1*-LR-DNAm signatures (11 of 44) and 73% of *ATRX* LR-DNAm signatures (19 of 26)] (Supplementary Tables 6, 7). A higher proportion of methylation changes at segmental duplications in ATR-X LR-DNAm might be related to the preferential localization and specific function of ATRX at repetitive elements, including pericentromeric heterochromatin [43, 44], as previously reported [36]. Notably, LR-DNAm could be evaluated in parallel with genetic variant discovery. Hence, it might be possible to perform a highly accurate diagnostic test combining genome and epigenome analyses using nanopore sequencing.

In terms of the disadvantages of LR-GS, first, read count-based scoring of DNA methylation level was employed in this study. In low-coverage datasets (Fig. 1C), accurate measurement of the DNA methylation level could not be achieved, suggesting that regional analysis after merging neighboring CpGs into a single region could be preferable for analysis, as in DMR detection in this study. Second, in nanopore sequencing, the quality of DNA has a significant impact on the sequencing results,

such as read length, depth of coverage, and phased block size (Fig. 1C, D). Such variable factors likely affect the number of DMRs detected in each analysis, leading to significant variability between samples. Finally, poor scalability is a significant obstacle. This resulted in insufficient statistical power and many false positive calls. We note that reported DNAm signature is based on the DNA methylation scoring of a set of single CpG sites, whereas LR-DNAm signature is the region-based scoring (i.e., DMR). Considering the forementioned disadvantage of long-read sequencing and different methodological principles between microarray and long-read sequencing, region-based alternative sets of DNAm (LR-DNAm signature) might be beneficial for the robust classification of DNAm signature. Ongoing improvements of long-read sequencing technologies with respect to the coverage, cost, and sequencing accuracy may overcome these obstacles.

We unexpectedly identified hypomethylation of a part *NSD1* CpGi. Hence, we screened the differential methylation at *NSD1* CpGi using conventional PCR-based assays (COBRA). This screening indeed detected statistically significant differences in methylation between Sotos syndrome and healthy controls. Importantly, this observation was confirmed in both cases of *NSD1* point mutations and those of a 5q35 submicroscopic deletion. In general, the applied fluorescence-based genetic testing method varies depending on the type and size of mutations, that is, fluorescence in situ hybridization (FISH) or microarray comparative genome hybridization (aCGH) for 5q35 submicroscopic deletions, and NGS or Sanger sequencing for point mutations. We showed that hypomethylation at *NSD1* CpGi was observed even in the wild-type (non-deleted) allele in cases of a 5q35 submicroscopic deletion (Fig. 5B), suggesting the advantage of independence from the type or size of pathogenic variants (Fig. 5C). We also showed that this DNA methylation test could be used to discriminate Sotos syndrome from other clinically overlapping OGIDs with 100% specificity, which is consistent with previous reports using multi-locus DNAm signature [37] (Fig. 5D). Hence, we suggest that *NSD1* CpGi is a good diagnostic marker for Sotos syndrome and could be simply investigated by a single-locus PCR-based assay in a rapid and inexpensive manner as a first-tier diagnostic assay.

We note the limitations of this study. The number of biological replicates/samples is usually limited due to cost constraints (expensive costs) in LRS [10]. To overcome this limitation at least in part, we used the DSS program as a tool to identify genomic regions with differential methylation rather than single probe in this study. An advantage of DSS is that smoothing the neighboring CpG sites can be viewed as pseudo-replicates, and find the differentially methylated regions rather than single base methylation changes. This can reduce the artifact (false positive) and help to identify DNAm signature with the reasonable precision [19]. In fact, our LR-based DNAm signatures can be successfully validated in independent validation cohort. At this time, relatively higher costs of LRS than the current gold-standard method, such as EPIC methylation array, limit the size of cohort. As the costs continue to decrease, independent evaluation of LR-DNAm signature with a large sample size will be needed, as was the case with follow-up study of EPIC array-based epigignature [9].

Conclusions

Distinct sets of differential methylation marks that are unique to Sotos and ATR-X syndromes, named LR-DNAm signatures, were found in this study by using long-read sequencing. We suggest that the advantages of long-read sequencing may enhance the detection of previously unidentified differential methylation marks, including valuable diagnostic markers, even in single-locus tests. This study may serve as a prototype for future genetic tests through simultaneously profiling genomic and epigenomic alterations using long-read sequencing technologies.

Abbreviations

ONT	Oxford Nanopore Technologies
PacBio	Pacific bioscience
CCS	Circular consensus sequencing
5-mC	5-Methylcytosine
DMR	Differentially methylated region
SVM	Support vector machine
ATR-X syndrome	Alpha thalassemia/mental retardation X-linked syndrome
COBRA	Combined bisulfite restriction analysis

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13148-025-01832-0.

Additional file1

Additional file2

Acknowledgements

We thank N. Watanabe, T. Miyama, M. Sato, S. Sugimoto, and K. Takabe for technical assistance. We also thank Edanz (https://jp.edanz.com/ac) for editing a draft of this manuscript.

Author contributions

T.M. and N.M. designed the study. T.M., T.H., N.N., M.S., L.F., Y.U., N.T., K.H., E.K., A.F., K.M., and S.M. contributed to acquisition and analysis of nanopore and PacBio long-read sequencing data. K.N. contributed to methylation array analyses. N.O. contributed to the clinical data analysis. T.M. wrote the original draft. T.M. and N.M. reviewed and edited the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by AMED under grant numbers JP24ek0109674, JP24ek0109760, JP24ek0109617, JP24ek0109648, and JP24ek0109677

(N. Matsumoto); JSPS KAKENHI under grant numbers JP23K27568 and JP23K18278 (T. Mizuguchi), JP24K18862 (M. Sakamoto), JP23K07229 (Y. Uchiyama), JP23K15353 (N. Tsuchida), JP21K07869 (E. Koshimizu), JP23K24490 (K. Nakabayashi), JP23K27520 (S. Miyatake), and JP24K02230 (N. Matsumoto); and the Takeda Science Foundation (T. Mizuguchi, N. Matsumoto).

Availability of data and materials
No datasets were generated or analyzed during the current study.

Declarations

Ethical approval and consent to participate
This study was approved by the Institutional Review Board of Yokohama City University School of Medicine. Written informed consent for inclusion in the study was obtained from all participants or their legal guardians.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Human Genetics, Graduate School of Medicine, Yokohama City University, 3-9 Fukuura, Kanazawa-ku, Yokohama 236-0004, Japan. [2]Department of Medical Genetics, Osaka Women's and Children's Hospital, Izumi, Japan. [3]Department of Rare Disease Genomics, Yokohama City University Hospital, Yokohama, Japan. [4]Department of Maternal-Fetal Biology, National Center for Child Health and Development, Tokyo, Japan. [5]Department of Clinical Genetics, Yokohama City University Hospital, Yokohama, Japan.

Received: 23 August 2024 Accepted: 4 February 2025
Published online: 18 February 2025

References

1. Feinberg AP. The key role of epigenetics in human disease prevention and mitigation. N Engl J Med. 2018;378:1323–34.
2. Dor Y, Cedar H. Principles of DNA methylation and their implications for biology and medicine. Lancet. 2018;392:777–86.
3. Chater-Diehl E, Goodman SJ, Cytrynbaum C, Turinsky AL, Choufani S, Weksberg R. Anatomy of DNA methylation signatures: emerging insights and applications. Am J Hum Genet. 2021;108:1359–66.
4. Haghshenas S, Bhai P, Aref-Eshghi E, Sadikovic B. Diagnostic utility of genome-wide DNA methylation analysis in Mendelian neurodevelopmental disorders. Int J Mol Sci. 2020;21(23):9303.
5. Turinsky AL, Choufani S, Lu K, Liu D, Mashouri P, Min D, et al. EpigenCentral: portal for DNA methylation data analysis and classification in rare diseases. Hum Mutat. 2020;41:1722–33.
6. Aref-Eshghi E, Kerkhof J, Pedro VP, Groupe DIF, Barat-Houari M, Ruiz-Pallares N, et al. Evaluation of DNA methylation episignatures for diagnosis and phenotype correlations in 42 Mendelian neurodevelopmental disorders. Am J Hum Genet. 2020;106:356–70.
7. Aref-Eshghi E, Bend EG, Colaiacovo S, Caudle M, Chakrabarti R, Napier M, et al. Diagnostic utility of genome-wide DNA methylation testing in genetically unsolved individuals with suspected hereditary conditions. Am J Hum Genet. 2019;104:685–700.
8. Aref-Eshghi E, Rodenhiser DI, Schenkel LC, Lin H, Skinner C, Ainsworth P, et al. Genomic DNA methylation signatures enable concurrent diagnosis and clinical genetic variant classification in neurodevelopmental syndromes. Am J Hum Genet. 2018;102:156–74.
9. Husson T, Lecoquierre F, Nicolas G, Richard AC, Afenjar A, Audebert-Bellanger S, et al. Episignatures in practice: independent evaluation of published episignatures for the molecular diagnostics of ten neurodevelopmental disorders. Eur J Hum Genet. 2024;32(2):190–9.
10. Mastrorosa FK, Miller DE, Eichler EE. Applications of long-read sequencing to Mendelian genetics. Genome Med. 2023;15:42.
11. Tse OYO, Jiang P, Cheng SH, Peng W, Shang H, Wong J, et al. Genome-wide detection of cytosine methylation by single molecule real-time sequencing. Proc Natl Acad Sci USA. 2021;118(5):e2019768118.
12. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. Detecting DNA cytosine methylation using nanopore sequencing. Nat Methods. 2017;14:407–10.
13. Fukuda H, Yamaguchi D, Nyquist K, Yabuki Y, Miyatake S, Uchiyama Y, 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;13:204.
14. Deng J, Zhou B, Yu J, Han X, Fu J, Li X, et al. Genetic origin of sporadic cases and RNA toxicity in neuronal intranuclear inclusion disease. J Med Genet. 2022;59(5):462–9.
15. Deng J, Yu J, Li P, Luan X, Cao L, Zhao J, et al. Expansion of GGC repeat in *GIPC1* is associated with oculopharyngodistal myopathy. Am J Hum Genet. 2020;106:793–804.
16. Li H. New strategies to improve minimap2 alignment accuracy. Bioinformatics. 2021;37:4572–4.
17. Shafin K, Pesout T, Chang PC, Nattestad M, Kolesnikov A, Goel S, et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. Nat Methods. 2021;18:1322–32.
18. Holt JM, Saunders CT, Rowell WJ, Kronenberg Z, Wenger AM, Eberle M. HiPhase: jointly phasing small, structural, and tandem repeat variants from HiFi sequencing. Bioinformatics. 2024;40(2):btae042.
19. Wu H, Xu T, Feng H, Chen L, Li B, Yao B, et al. Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. Nucleic Acids Res. 2015;43:e141.
20. Court F, Tayama C, Romanelli V, Martin-Trujillo A, Iglesias-Platas I, Okamura K, et al. Genome-wide parent-of-origin DNA methylation analysis reveals the intricacies of human imprinting and suggests a germline methylation-independent mechanism of establishment. Genome Res. 2014;24:554–69.
21. Liu Y, Rosikiewicz W, Pan Z, Jillette N, Wang P, Taghbalout A, et al. DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. Genome Biol. 2021;22:295.
22. Akbari V, Garant JM, O'Neill K, Pandoh P, Moore R, Marra MA, et al. Megabase-scale methylation phasing using nanopore long reads and NanoMethPhase. Genome Biol. 2021;22:68.
23. Morgan R, Loh E, Singh D, Mendizabal I, Yi SV. DNA methylation differences between the female and male X chromosomes in human brain. bioRxiv. 2024. https://doi.org/10.1101/2024.04.16.589778.
24. Horsthemke B. Mechanisms of imprint dysregulation. Am J Med Genet C Semin Med Genet. 2010;154C:321–8.
25. Zink F, Magnusdottir DN, Magnusson OT, Walker NJ, Morris TJ, Sigurdsson A, et al. Insights into imprinting from parent-of-origin phased methylomes and transcriptomes. Nat Genet. 2018;50:1542–52.
26. Hernandez Mora JR, Tayama C, Sanchez-Delgado M, Monteagudo-Sanchez A, Hata K, Ogata T, et al. Characterization of parent-of-origin methylation using the Illumina Infinium MethylationEPIC array platform. Epigenomics. 2018;10:941–54.
27. Joshi RS, Garg P, Zaitlen N, Lappalainen T, Watson CT, Azam N, et al. DNA methylation profiling of uniparental disomy subjects provides a map of parental epigenetic bias in the human genome. Am J Hum Genet. 2016;99:555–66.
28. Cerrato F, Sparago A, Ariani F, Brugnoletti F, Calzari L, Coppede F, et al. DNA methylation in the diagnosis of monogenic diseases. Genes (Basel). 2020;11(4):355.
29. Akbari V, Hanlon VCT, O'Neill K, Lefebvre L, Schrader KA, Lansdorp PM, et al. Parent-of-origin detection and chromosome-scale haplotyping using long-read DNA methylation sequencing and Strand-seq. Cell Genom. 2023;3:100233.
30. Loyfer N, Magenheim J, Peretz A, Cann G, Bredno J, Klochendler A, et al. A DNA methylation atlas of normal human cell types. Nature. 2023;613:355–64.
31. Soejima H, Higashimoto K. Epigenetic and genetic alterations of the imprinting disorder Beckwith-Wiedemann syndrome and related disorders. J Hum Genet. 2013;58:402–9.
32. Mandelker D, Schmidt RJ, Ankala A, McDonald Gibson K, Bowser M, Sharma H, et al. Navigating highly homologous genes in a molecular

diagnostic setting: a resource for clinical next-generation sequencing. Genet Med. 2016;18:1282–9.
33. Sone J, Mitsuhashi S, Fujita A, Mizuguchi T, Hamanaka K, Mori K, et al. Long-read sequencing identifies GGC repeat expansions in *NOTCH2NLC* associated with neuronal intranuclear inclusion disease. Nat Genet. 2019;51:1215–21.
34. Ishiura H, Shibata S, Yoshimura J, Suzuki Y, Qu W, Doi K, et al. Noncoding CGG repeat expansions in neuronal intranuclear inclusion disease, oculopharyngodistal myopathy and an overlapping disease. Nat Genet. 2019;51:1222–32.
35. Liufu T, Zheng Y, Yu J, Yuan Y, Wang Z, Deng J, et al. The polyG diseases: a new disease entity. Acta Neuropathol Commun. 2022;10:79.
36. Schenkel LC, Kernohan KD, McBride A, Reina D, Hodge A, Ainsworth PJ, et al. Identification of epigenetic signature associated with alpha thalassemia/mental retardation X-linked syndrome. Epigenetics Chromatin. 2017;10:10.
37. Choufani S, Cytrynbaum C, Chung BH, Turinsky AL, Grafodatskaya D, Chen YA, et al. *NSD1* mutations generate a genome-wide DNA methylation signature. Nat Commun. 2015;6:10207.
38. Conteduca G, Testa B, Baldo C, Arado A, Malacarne M, Candiano G, et al. Identification of alternative transcripts of *NSD1* gene in Sotos Syndrome patients and healthy subjects. Gene. 2023;851: 146970.
39. Bilichak A, Kovalchuk I. The combined bisulfite restriction analysis (COBRA) assay for the analysis of locus-specific changes in methylation patterns. Methods Mol Biol. 2017;1456:63–71.
40. Kurotaki N, Imaizumi K, Harada N, Masuno M, Kondoh T, Nagai T, et al. Haploinsufficiency of *NSD1* causes Sotos syndrome. Nat Genet. 2002;30:365–6.
41. Choufani S, Gibson WT, Turinsky AL, Chung BHY, Wang T, Garg K, et al. DNA methylation signature for *EZH2* functionally classifies sequence variants in three PRC2 complex genes. Am J Hum Genet. 2020;106:596–610.
42. Barros-Silva D, Marques CJ, Henrique R, Jeronimo C. Profiling DNA methylation based on next-generation sequencing approaches: New insights and clinical applications. Genes (Basel). 2018;9(9):429.
43. Law MJ, Lower KM, Voon HP, Hughes JR, Garrick D, Viprakasit V, et al. ATR-X syndrome protein targets tandem repeats and influences allele-specific expression in a size-dependent manner. Cell. 2010;143:367–78.
44. McDowell TL, Gibbons RJ, Sutherland H, O'Rourke DM, Bickmore WA, Pombo A, et al. Localization of a putative transcriptional regulator (ATRX) at pericentromeric heterochromatin and the short arms of acrocentric chromosomes. Proc Natl Acad Sci USA. 1999;96:13983–8.

Publisher's Note