Genome Medicine

RESEARCH

Open Access

Genome-wide methylation detection and episignature analysis using PacBio long-read sequencing

Véronique Ivashchenko[1,2], Michelle de Groot[1], Ronny Derks[1,3], Amber den Ouden[1], Gelana Khazeeva[1,3], Simone van den Heuvel[1], Raoul Timmermans[1], Jordi Corominas Galbany[1], Rolph Pfundt[1], Tom Hofste[1], Helger Yntema[1], Lisenka Vissers[1,3], Alexander Hoischen[1,3,4], Juliet Hampstead[1,3†] and Christian Gilissen[1,3*†]

Abstract

Background The detection of 5-methylcytosine (5mC) patterns in the human genome is relevant for the diagnosis of various genetic conditions. Genome-wide methylation episignatures provide a new approach for resolving variants of uncertain significance. Whole-genome DNA methylation detection is usually performed using short-read bisulfite sequencing procedures or methylation arrays. Pacific Biosciences (PacBio) long-read sequencing (LRS) is a technology that can detect 5mC on native DNA. Here we assessed the ability of PacBio HiFi LRS to robustly identify genome-wide methylation for medical purposes. We assessed the ability to identify differentially methylated imprinted regions and genome-wide epigenetic signatures for genetic disorders.

Methods PacBio LRS methylation detection was compared to two conventional genome-wide DNA methylation sequencing techniques. We used 30 PacBio LRS samples (10 trios at ~ 30× coverage) and assessed the ability to detect differential methylation using 25 known parentally imprinted regions with different genome-wide sequence coverage. Finally, we evaluated two published epigenetic signatures for *KMT2A* gene defects using PacBio LRS data from 12 *KMT2A* patients, 2 patients with a variant of uncertain significance, and 39 control samples.

Results PacBio methylation calls on the HG002 cell line were highly comparable to short-read protocols, providing 5% additional calls, predominantly in regions that are difficult to assess using short-read technologies. Among the 10 trios, correct haplotype-resolved methylation patterns were found for 24 out of 25 (96%) imprinted regions and all of the methylated alleles showed the expected parental origin. Downsampling analysis showed that these imprinted regions can robustly be detected using a minimum of 15× genome-wide coverage. Finally, we found that although the two published *KMT2A* episignatures are significantly different, both are successfully able to distinguish *KMT2A* patients from controls and classify the two *KMT2A* VUS samples as control samples.

Conclusions Overall, our results indicate that PacBio LRS can reliably detect specific medically relevant methylation changes as well as genome-wide episignatures in rare disease patients.

Keywords Methylation, Episignature, Genomic imprinting, PacBio long-read sequencing

†Juliet Hampstead and Christian Gilissen contributed equally to this work.

*Correspondence:
Christian Gilissen
christian.gilissen@radboudumc.nl
Full list of author information is available at the end of the article

Background

About 70–80% of the human genome is subject to 5-methylcytosine (5mC) DNA methylation [1]. DNA methylation patterns are cell-type specific, and several loci have been shown to vary among individuals based on different factors such as age, sex, or ethnicity [2]. 5mC methylation can also be allele specific, which can lead to differential methylation of CpG sites. This methylation pattern is especially found in imprinting control regions (ICR), which are germline differentially methylated regions (DMRs) used to control the expression of imprinted genes [3]. Differences in DNA methylation have been linked to various genetic disorders such as Beckwith-Wiedemann syndrome, Prader-Willi syndrome, and Angelman syndrome [4]. These imprinting disorders result from genetic aberrations affecting imprinted genes, which have mono-allelic expression in a parent-of-origin manner (either from maternal or paternal gene copy) [3]. Some repeat-instability disorders can also drive DNA methylation defects, such as fragile X syndrome (FXS) [5]. Other genetic disorders can involve mutations of key proteins underlying DNA methylation processes such as methyl binding proteins (i.e., *MECP2*) in the case of Rett syndrome [6]. In addition, a growing number of Mendelian disorders have also been associated with specific peripheral blood DNA methylation profiles called "episignatures." These signatures are based on the definition of hundreds of differentially methylated CpG sites between case and control samples. Such episignatures can be used as biomarkers to guide diagnosis for rare disease patients with nonspecific or heterogeneous clinical presentations, as well as complementary functional genomic tests to support re-classification of variants of uncertain significance (VUS) revealed by previous genetic screenings [7, 8].

Identification of genome-wide 5mC epigenetic marks commonly relies on bisulfite-based short-read sequencing approaches. However, bisulfite treatment may partially degrade DNA and conversion of unmethylated cytosines to thymines induces a loss of diversity of sequenced reads which can cause challenges during alignment [9]. Methylation detection on repeated sequences of DNA such as CpG islands might also be limiting for short-read sequencing methods [10]. The majority of published episignatures are generated through the use of methylation arrays [11]. Arrays are more affordable than genome-wide bisulfite sequencing which explains their use in large-scale case–control comparison studies. Yet, arrays are limited to the screening of a preselected subset of CpG sites genome-wide and share the limitations of bisulfite treatment of DNA [12, 13]. In the past few years, long-read sequencing (LRS) techniques have emerged that generate sequencing reads > 15 kb but also natively detect 5mC methylation without the need of extra DNA treatment steps [14, 15]. Both PacBio and Oxford Nanopore Technologies (ONT) LRS platforms have been shown to report per CpG methylation values consistent with those obtained with methylation arrays and short-read bisulfite sequencing strategies [16–18]. In the case of PacBio sequencing, methylated bases are detected during real-time sequencing through their effect on polymerase kinetic parameters, such as pulse width (PW) and inter-pulse duration (IPD, referring to the interval of time between two successive fluorescent pulses) [14].

Previous studies have assessed the general ability and reliability of methylation detection in human samples using PacBio long-read sequencing. Foox et al. compared methylation detection of multiple short-read sequencing strategies and showed that enzymatic-based EMSeq and bisulfite-based MethylSeq were highly comparable top performing methods [19]. Ni et al. compared methylation detection from PacBio HiFi to bisulfite sequencing for multiple depth of coverage thresholds (5× to 25× coverage) and reported a Pearson correlation between the two detection sets close to 0.9 at 20× coverage [16].

Here, we assessed the ability of PacBio HiFi LRS to identify genome-wide 5mC methylation for medical purposes. First, we compared the overall methylation detection provided by PacBio LRS platform with other conventional genome-wide DNA methylation sequencing techniques using public data from the Genome In A Bottle (GIAB) HG002 cell line [20]. Next, we used LRS data from 10 in-house proband-parent trios to assess the ability of PacBio to detect methylation in 25 known parentally imprinted regions, of which 13 are associated with an imprinting disorder. Finally, we investigated the ability of PacBio sequencing to detect genome-wide epigenetic signatures for genetic disorders using 12 patients with Wiedemann-Steiner syndrome (WSS) with known *KMT2A* pathogenic variants and 39 control samples.

Methods

Methylation sequencing datasets

HG002 PacBio long-read sequencing data

Methylation detection was first assessed using publicly available PacBio data containing aligned HiFi long reads with 5mC tags from the online PacBio repository [21] (Additional file 1: Table S1). Sequencing data is derived from the HG002/NA24385 sample, which is an EBV-transformed lymphoblastoid cell line [20]. Methods used for the processing of this sample are described on the repository webpage. Downstream analysis included HiFi reads generation with CCS [22] (version 6.0.0) with "--all-kinetics" option to generate consensus kinetics tags, CpG methylation annotation with primrose [23] (version 1.0.0),

alignment to the GRCh38/Hg38 genome with pbmm2 [24] (version 1.4.0), small variant (substitution and indel) calling with DeepVariant [25] (version 1.0.0), and phasing with WhatsHap [26] (version 1.0). We used this data as input for pb-cpg-tools [27] (version 2.3.1) with options "--pileup-mode model --modsites-mode denovo" for the de novo detection mode and "--pileup-mode model --modsites-mode reference" for the reference detection mode, to generate corresponding HG002 PacBio BED files.

HG002 short-read sequencing data from the EpiQC study

Methylation calls from PacBio HiFi LRS were compared with two other whole-genome DNA methylation sequencing protocols selected from the EpiQC study [19] based on report files available for the HG002 cell line on the EpiQC study repository [28]. Two short-read assays for whole-genome DNA methylation detection were selected: Enzymatic Methyl-Seq (EMSeq), an enzymatic deamination protocol, and Swift Bio sciences Accel-NGS Methyl-Seq (MethylSeq), a whole-genome bisulfite sequencing protocol (Additional file 1: Table S1). These two methods were chosen among the best performing protocols in the study. In these two protocols, sequenced reads were aligned on GRCh38 human reference genome using bwa-meth [29] (version 0.2.1), and methylation called with MethylDackel [30] extract with "--mergeContext" parameter to merge calls per strand into dinucleotide context. Methylation calls were stored with 0-based genomic coordinates in report files with bedGraph format, further downloaded from the study repository. According to methods performed in the EpiQC study, we combined all technical replicates per library type per cell line per laboratory by summing the methylated and unmethylated counts per CpG site. MethylSeq and EMSeq data were downsampled to a 30× mean depth of coverage (as described below), to match the depth of coverage in the combined HG002 PacBio BED file.

Proband/parent LRS trio cohort

Methylation detection on PacBio HiFi long reads was further assessed on 10 in-house trios (30 whole-blood samples in total), as summarized in Additional file 1: Table S1. Probands of these trios had an unresolved neurodevelopmental disorder with no disease-causing variant identified from previous sequencing studies. The DNA isolation was done using the Chemagic DNA Blood 4 K kit (PerkinElmer) on the Chemagic Star to extract high molecular weight (HMW) DNA from archived blood samples. Whole-genome LRS of these samples was performed on the PacBio Revio® instrument. The targeted depth of coverage for HiFi reads was 30× (Additional file 1: Table S2). All samples were processed the same way according to manufacturer's instructions (PacBio, Menlo Park, CA, USA). DNA shearing was performed using Megaruptor 3 (Diagenode, Liège, Belgium) on 7 µg DNA, to a target size of 15–18 kb. Library preparation relied on SMRTbell express template prep kit 3.0 (PacBio, Menlo Park, CA, USA). Libraries were size-selected > 10 kb on the BluePippin (Sage Science, Beverly, MA, USA). Libraries were further sequenced for 24 h on the Revio® instrument (ICS 12.0.4). HiFi reads were processed with an in-house script which is available on Github [31], including alignment to the GRCh38 reference genome (pbmm2 [24] version 1.13.1). Structural variants (PBSV [32] version 2.9.0) and small variants (DeepVariant [25] version 1.5.0) were identified after phasing (HiPhase [33] version 1.1.0). Methylation calls were generated (pb-cpg-tools [27] version 2.3.1).

KMT2A LRS cohort for episignature analysis

KMT2A genome-wide episignatures were investigated based on a cohort of 12 patients with known *KMT2A* loss-of-function mutations (cases) and 39 control samples (case to control ratio of 1:3). Cases and controls were selected from our in-house clinical cohort based on previous genetic testing results. Loss-of-function variants included 5 stops, 4 frameshifts, and 3 splice site variants classified as pathogenic (ACMG class 5) or probably pathogenic (ACMG class 4). Two samples with *KMT2A* missense variants of uncertain significance (VUS) variants (ACMG class 3) were also included for analysis (Additional file 1: Tables S1, S4). Variant consequences were reported according to the NM_001197104.2 MANE Select RefSeq *KMT2A* transcript. Control samples were selected among in-house patients with resolved genetic disorders that are not associated with methylation defects. Case and control groups were balanced for males and females. Case samples consisted of 6 males and 6 females individuals. Control samples consisted of 21 males and 18 females. Patients with *KMT2A* VUS variants consisted of 1 male and 1 female individual.

Whole-genome LRS of these samples was performed on the PacBio Revio® instrument, processed the same as previously described for our trio cohort. These samples were sequenced based on 12 different sequencing batches (9 sequencing batches for control samples and 3 sequencing batches for *KMT2A* samples). We did not apply any batch correction to the methylation data reported by PacBio sequencing. The targeted depth of coverage for HiFi reads was 15× for case samples and 30× for controls. Case samples were sequenced at 15× by processing two samples per SMRT cell to reduce sequencing costs. Control samples, originally sequenced for other studies

at 30× coverage (one sample per SMRT cell), were downsampled from aligned BAM files using samtools [34] (version 1.11) to a mean depth of coverage of 15× in order to match the cases. New methylation probabilities were generated on these downsampled files with pb-cpg-tools [27] (version 2.3.1) with options "--pileup-mode model --modsites-mode denovo".

External datasets

Reference genome annotations

The annotations of CpG islands, RefSeq Functional Elements, centromeric regions, and ENCODE blacklist regions were downloaded from corresponding tracks of the UCSC Genome Browser [35].

We selected 25 known DMRs associated with imprinted genes reported by Joshi et al [36]. In order to relate DMR identifier to corresponding genomic location, we refer to selected DMRs using the format: "DMR number (closest RefSeq Gene or Gene promoter)" (Additional file 1: Table S3). These DMRs were delineated from comparisons of uniparental disomy patients to biparental control subjects. We chose to restrain our analysis to the DMRs based on two criteria: (1) selected regions had to contain at least 5 probes on a 450 k array, (2) selected regions had a mean methylation in peripheral blood between 40 and 60% across a cohort of 4004 individuals from the general population [36]. To provide a more comprehensive comparison of probe coverage, we included probe coverage data from a recent study by Jima et al. which used WGBS to identify imprinted control regions (ICRs) [37].

Age-related CpG sites from literature

Methylation patterns are known to change with age [38], which can introduce confounders when comparing cases and controls. To avoid this, we needed to ensure there was no significant age difference between the two groups. However, age information for both case and control samples was unavailable. Alternatively, we selected three sets of age-related CpG sites from the literature, based on methylation array studies, and used them to assess potential age differences between *KMT2A* cases and controls.

Hannum et al. CpGs dataset consisted of 71 CpG methylation markers reported as highly predictive of age, based on analysis of 656 human individuals whole blood using the Illumina Infinium HumanMethylation450 BeadChip assay [39]. Array identifiers of corresponding CpG markers were downloaded from the publication repository. Genomic locations of these markers on the GRCh38 genome assembly were extracted using the EPIC array reference BED file from the UCSC Genome Browser [37]. We were able to successfully convert 65 out of 71 cg array markers to GRCh38 coordinates.

Horvath et al. CpGs dataset consisted of 353 CpG age-related methylation markers, based on the analysis of 8000 samples from 51 human healthy tissues and cell types using Illumina 27 K and Illumina 450 K array platforms [40]. Array identifiers of corresponding CpG markers were downloaded from the publication repository. Genomic locations of these markers on the GRCh38 genome assembly were extracted using the EPIC array reference BED file from the UCSC Genome Browser. We were able to successfully convert 334 out of 353 cg array markers to GRCh38 coordinates.

Levine et al. CpGs dataset consisted of 513 CpG methylation markers reported as predictive of phenotypic age, based on the analysis of 456 human individuals whole-blood methylation array data restricted to CpG sites common to Illumina 27 K, 450 K, and EPIC array platforms [41]. Array identifiers of corresponding CpG markers were downloaded from the publication repository. Genomic locations of these markers on the GRCh38 genome assembly were extracted using the EPIC array reference BED file from the UCSC Genome Browser. We were able to successfully convert all the 513 cg array markers to GRCh38 coordinates.

Array-based KMT2A episignatures

Two sets of published *KMT2A* signature sites were selected from Foroutan et al. [42] and Hampstead [43] publications. Both of these episignatures were generated using case–control comparisons on methylation array data.

The Foroutan et al. training set consisted of 41 cases with pathogenic *KMT2A* variants (20 nonsense, 14 frameshift, 4 missense, 3 splice site variants) and 82 age, sex, and array type-matched controls (case to control ratio of 1:2) [42]. The corresponding DNA methylation signature included 207 differentially methylated probes genome-wide. A BED file containing hg19 genomic locations of the differentially methylated CpG sites and the case–control methylation difference per site was downloaded from the publication repository. Genomic locations were further converted to GRCh38 genome assembly using UCSC LiftOver [35].

The Hampstead training set consisted of 16 cases with clinically interpreted pathogenic/likely pathogenic loss-of-function variants in *KMT2A* and 37 age and sex-matched negative controls (case to control ratio of 1:2) [43]. The corresponding DNA methylation signature included 879 differentially methylated probes genome-wide. A BED file containing GRCh38 genomic locations of the differentially methylated CpG sites and the mean

methylation for case and control groups per site was obtained on request from the author.

Statistical methods

Downsampling of methylation calls

Downsampling of methylation calls from MethylSeq and EMSeq protocols for HG002 cell line were performed from bedGraph files with downsampling script from EpiQC study (available at https://github.com/nebiolabs/methylation_tools/blob/master/downsample_methylKit.py). In this script, the fraction of read counts kept in each file was set to match the desired downsampling depth of coverage. Downsampling from bedGraph files has been shown to give similar results in terms of read counts, number of CpG sites detected, and methylation calls as the downsampling from BAM files, while being less demanding in disk space and computing time [19]. Contrary to this, downsampling of methylation calls from PacBio LRS cannot be performed directly on BED files generated with pb-cpg-tools [27]. Methylation probabilities reported in these files do not correspond to raw ratio of methylated over unmethylated numbers of reads (as it is in bedGraph files) but have been corrected by the intermediate use of a model. Considering this, the downsampling of PacBio LRS samples has been done from aligned BAM files using samtools [34] (version 1.11).

To assess the impact of sequencing depth on the methylation detection from PacBio LRS samples, we downsampled data from 30× to 5× in 14 samples from the trio cohort with initial coverage ≥ 30× using samtools [34] (version 1.18). Methylation calls at each downsampled threshold were computed using pb-cpg-tools [27] (version 2.3.2) with options "--pileup-mode model --modsites-mode denovo". The methylation calls from downsampled samples (25×, 20×, 15×, 10×, and 5×) were compared to the 30× reference. We focused on CpG sites where methylation probabilities fell within ± 1 standard deviation (SD) of the 30× median to exclude highly variable sites. The proportion of CpG sites overlapping the 30× reference was calculated at each depth. Pearson correlation was calculated using the scipy stats package [44] and used to compare CpG methylation probabilities between 30× and lower depths. Mean absolute error (MAE) was computed to quantify deviations in methylation probabilities between downsampled and 30× data.

Phasing analysis Trio sequencing enabled us to investigate the parental origin of methylated allele in proband by comparing the variants exclusively shared by the proband and one of his parents. This was done manually for the ten in-house trios using IGV [45] and compared to the expected parental origin of methylation from literature [36].

Differential methylation analysis on PacBio LRS data We built an analytical algorithm to uncover differentially methylated CpG sites between two training groups of PacBio LRS samples (Additional file 1: Table S5). In order to show that PacBio methylation can be used for differential methylation discovery, we first applied this strategy to compare methylation probabilities of 8 female and 29 male samples selected from the in-house episignature cohort (Additional file 1: Table S1). The female:male ratio (8:29) was set to approximate the 1:3 ratio of the *KMT2A* PacBio in-house case:control cohort (11:38). The differential methylation analysis was performed on chromosomes common to both male and female groups (chr1-22,X). Before analysis, a principal component analysis (PCA) was performed to ensure that male and female samples do not cluster separately prior comparison. Statistical analysis was preceded by two successive filtering steps. The first filter consisted in the selection of common CpG sites with minimal 6× depth of coverage in all female samples and at least 80% of the male samples. This step was set to filter out possible artifacts and poorly covered CpG sites where we expect methylation detection to be less accurate. Because of the millions of CpG sites reported genome-wide, a second filter was needed to restrict the number of CpG sites prior comparison. This filter consisted of a minimal difference threshold of 20% between the median methylation probabilities of male and female groups. Statistical analysis was based on cross-validation runs, with one sample left out of the training and used as a testing sample per run. For each CpG site that passed initial filtering steps, a Mann–Whitney *U* test was performed on the methylation probabilities of the two training groups. The corresponding *p* values were corrected for multi testing using FDR correction. CpG sites with adjusted *p* values under 0.05 threshold were considered to have a statistically significant differential methylation per run. Differentially methylated CpG sites extracted from the consensus of all the cross-validation runs were considered as the DNA methylation signature.

Power analysis To estimate power, we created Gaussian probability distributions for each CpG site covered ≥ 6× in PacBio data on the Illumina EPIC array. We used only probes covered by Illumina's EPIC methylation array because several studies have published *KMT2A* methylation signatures using samples methylation profiled on array [42, 43], allowing us to know a priori whether or not these CpG sites are included in the *KMT2A* signature. The mean and standard deviation of the Gaussian distribution for each site were derived from *KMT2A* case samples (case distribution) and control samples (control distribution). Methylation probabilities for simulated samples, from 10 to 100, were randomly drawn from these Gauss-

ian distributions at EPIC array CpG sites to simulate cases and controls. Power was calculated empirically by looking at the 75 array signature sites overlapping between both signatures and determining the difference between the number of signature CpG sites that are significant at an FDR-corrected Mann–Whitney *U p* value of 0.05 (true positives) and the number of EPIC array CpG sites not in methylation signatures that are significant at an FDR-corrected Mann–Whitney *U p* value of 0.05 (false positives).

Results

Using Genome In A Bottle data for the HG002 sample, we first compared methylation detection between PacBio LRS and two short-read whole-genome DNA methylation sequencing assays selected from the EpiQC study [19]: MethylSeq (bisulfite) and EMSeq (enzymatic). Overall, the three platforms shared ~ 93% of all reported CpG positions with a Pearson correlation > 0.90 for all pairwise comparisons (Additional file 1: Fig. S1A), similar to what was reported previously [16]. Because most CpG sites in the human genome are methylated, we additionally tested whether these correlations could have possibly occurred by chance. To test this, we created observed and permuted distributions at 10,000 randomly selected human CpG sites. We observed that the Spearman correlation coefficients of the observed distribution was significantly higher than that of the permuted distribution for both PacBio—EMSeq (Wilcoxon $p < 2.2 \times 10^{-20}$) and PacBio—MethylSeq (Wilcoxon $p < 2.2 \times 10^{-20}$) comparisons (Additional file 1: Fig. S2). In addition, we found that 696,375 (2.4%) of reported CpG sites were exclusive to the PacBio detection (Additional file 1: Fig. S1B). Out of these, 365,823 (52.5%) CpG sites overlapped with centromeric regions (Additional file 1: Fig. S3A) and 511,278 (73.4%) with the ENCODE blacklist regions (Additional file 1: Fig. S3B). These observations emphasize the ability of PacBio LRS to detect methylation on complex genomic regions that might be challenging for short-read methods.

PacBio methylation detection can be done in de novo and reference detection modes. The reference mode only reports CpG sites that are present in the reference genome, while de novo mode reports all CpG sites found in the consensus build from reads thus enabling the detection of new CpG sites created by genomic variants. We found that PacBio LRS de novo mode detected methylation for 1,320,727 (4.47%) additional CpG sites (Additional file 1: Fig. S1A). These CpG sites included 55,837 sites overlapping CpG islands (Additional file 1: Fig. S4A) and 1325 sites overlapping RefSeq Functional Elements (Additional file 1: Fig. S4B). Manual investigation of a subset of these sites using IGV showed that they indeed correspond to new CpG sites created by genomic variants (Additional file 1: Fig. S5). These results show that PacBio LRS enables the detection of new genomic CpG sites, which might be of particular interest for genomic variants occurring in regulatory genomic regions such as CpG islands or enhancer sequences, where they might modulate the overall epigenetic regional profile.

Validation of methylation detection on imprinted genomic regions

CpG methylation detection using HiFi LRS is accurate on in-house peripheral blood trio samples

We aimed to assess the ability and robustness of PacBio LRS to identify methylation changes for medically relevant differentially methylated regions in humans. For this we used data from 10 in-house proband-parent trios (30 samples) that were sequenced with a target sequencing depth of coverage of 30× on the PacBio Revio platform (Additional file 1: Tables S1, S2).

We investigated methylation detection on a set of 25 imprinted regions (22 maternal, 3 paternal) from literature including 13 regions linked to imprinting disorders in humans [4, 36] (Additional file 1: Table S3). Imprinted regions are expected to display a methylation level close to 50% because each allele is differentially methylated based on the parent of origin. As such these regions may serve as within sample control for detecting differential methylation. These 25 imprinted regions were considered as stable with a mean methylation level on 450 k array data on peripheral blood between 40 and 60% across a cohort of 4004 individuals from the general population [36]. This methylation profile was concordant on the 30 PacBio in-house peripheral blood samples (Fig. 1A) as for the HG002 cell-line and short-read sequencing data (Additional file 1: Fig. S6). We found that the number of CpG sites reported per imprinted region was five times higher with PacBio LRS data compared to the 450k array, and that PacBio LRS provided similar CpG coverage to WGBS data (Additional file 1: Table S3).

One advantage of LRS is that it allows for variant phasing and the separation of the two different alleles, based on which we can assess allele-specific methylation. For 24 out of 25 regions (96%), we found the expected differential methylation pattern between alleles across the majority of proband's peripheral blood samples (Fig. 1B), consisting of one allele harboring mean methylation > 60% (hypermethylated) and the other < 40% (hypomethylated). Phasing was challenging in a minority of probands and likely due to the lack of informative variants (Additional file 1: Fig. S7). The only exception was the imprinted region 16 (*SNRPN*) with half of the proband samples showing loss of differential methylation (Fig. 1B). We investigated methylation profiles on IGV for several of these samples (Additional file 1:

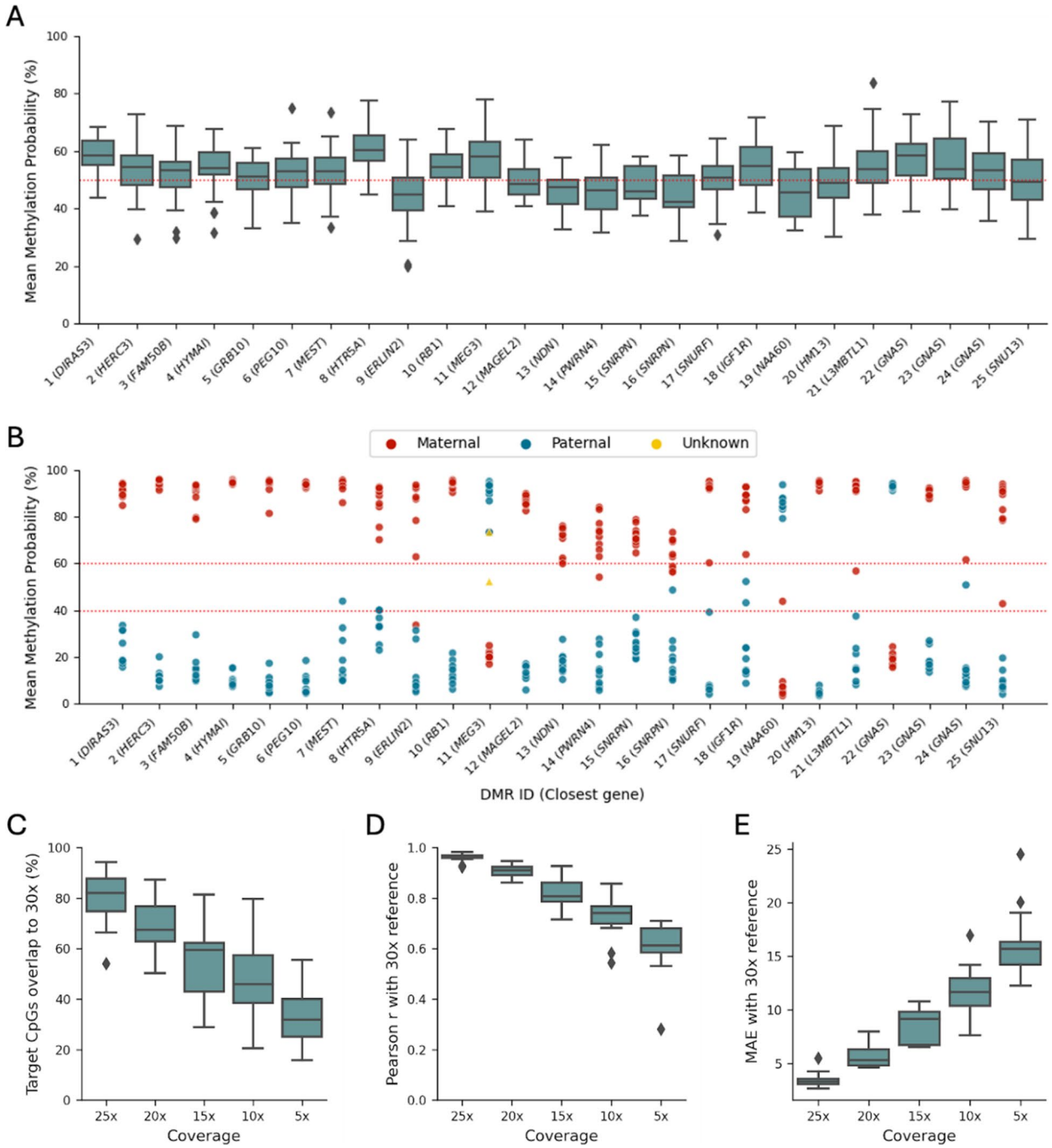

Fig. 1 Evaluation of PacBio LRS methylation detection on in-house trio samples for 25 stable imprinted genomic regions. **A** Mean methylation across 30 in-house trio samples for the 25 imprinted DMRs. **B** Methylation per haplotype in probands, with parental origin indicated. Parental origin is indicated for each haplotype. For one sample, the parental origin in region 11 could not be determined (see Supplementary Fig. S7A), which is labeled unknown. Impact of decreased sequencing depth on PacBio methylation detection (distributions for a subset of 14 in-house trio samples with > 30× initial coverage): **C** percentage of target CpGs overlapping the 30× reference at decreasing coverage thresholds, **D** Pearson correlation of CpG methylation probabilities at 30× versus lower depths, **E** mean absolute error (MAE) of CpG methylation probabilities between lower depths and 30× coverage

Fig. S8). Phasing appeared correct on IGV inspection and both alleles showed methylated and unmethylated CpGs, confirming absence of allele-specific methylation. The fact that our data include proband/parent trio samples also enabled us to investigate the parental origin of the methylated allele in the proband (Additional file 1: Fig. S9). For all the differentially methylated regions in all 10 probands, we identified a matching parental origin of the methylated allele in accordance with what has been described in the literature (Fig. 1B and Additional file 1: Table S3).

Overall, these results show that PacBio LRS methylation performs accurate methylation detection and haplotype-resolved methylation profiles on in-house peripheral blood samples.

Depth of coverage variation has an impact on the accuracy of methylation detection from HiFi LRS

Although our results show the reliability and potential advantages of whole-genome methylation detection using LRS, the cost for a 30× genome may still be prohibitive for large-scale application [46]. Therefore, we investigated the effect of sequencing depth on accurate methylation detection. We performed a downsampling procedure for LRS with minimal depths of coverage thresholds ranging from 30× to 5× for 14 out of the 30 in-house trio samples which had an initial sequencing depth of coverage over equal 30× (Additional file 1: Table S2). Then, we compared the downsampled methylation calls to those obtained with the 30× depth of coverage threshold on the 25 imprinted regions (DMRs). By restricting the analysis to CpG sites with methylation probabilities within ± 1 standard deviation (SD) of the 30× median, we excluded highly variable sites. As sequencing depth decreased, we observed a linear decline in accuracy. At 15×, the Pearson correlation with 30× remained around 0.8, with ~ 60% of CpG sites still overlapping (Fig. 1C, D). The mean absolute error (MAE) of the methylation probabilities ranged between 7 and 10 (Fig. 1E). While there is no clear cutoff for an optimal sequencing depth, 15× was a practical choice for our analysis, as it allows pooling two samples per SMRT cell while still providing consistent methylation probabilities.

Applicability of array-based *KMT2A* episignatures on PacBio HiFi LRS data

Episignatures represent the association of hundreds of differentially methylated CpG sites genome-wide, shown to have the better ability to distinguish case samples from controls. These signatures can further be used as biomarkers to guide the diagnosis for rare disease patients [7]. Genome-wide episignatures have been developed for several different genetic disorders based on DNA methylation arrays [47]. We chose the gene *KMT2A* based on the existence of published episignature CpG sites and the presence of patient samples in our center. First, we compared existing array-based *KMT2A* signatures and the applicability of these CpG sites to distinguish cases from controls on PacBio methylation data.

Array-based KMT2A signatures CpG sites can be used to differentiate case and control samples on PacBio HiFi data

KMT2A germline pathogenic variants are associated with the Wiedemann-Steiner syndrome (WSS, MIM# 605130). This is a rare autosomal dominant neurodevelopmental disorder for which the majority of reported pathogenic variants are de novo protein truncating variants [48]. We used two independent sets of published *KMT2A* signature sites that consisted of 207 (Foroutan et al. [42]) and 879 (Hampstead [43]) CpG sites genome-wide. Both of these signatures were extracted from cases/control comparisons using methylation arrays. First, we compared the CpG sites between the two array-based episignatures. We observed that these sites overlapped only partially on 75 CpG sites (7.42% of all CpG sites). Nevertheless, these common CpG sites displayed the same methylation difference between cases and controls, with ~ 90% of the sites showing a decrease of methylation in cases compared to controls (Fig. 2A). The 75 CpG sites common between the two signatures had a high correlation ($r = 0.9769$) of the cases vs controls methylation difference between the two array datasets (Additional file 1: Fig. S10).

We further investigated whether array-based *KMT2A* signatures can be used to differentiate samples with pathogenic *KMT2A* variants on PacBio LRS data. Based on our downsampling analysis and the practical advantage of combining two samples on a single SMRT cell, we performed 15× coverage sequencing on the PacBio Revio system for 12 patients with WSS with known (loss-of-function) pathogenic variants in *KMT2A* (cases) and 2 patients with a VUS in *KMT2A*. As controls, we used 39 samples from other studies, downsampled to 15× and sequenced on the same platform (Additional file 1: Tables S1, S4). Because these samples were distributed across different sequencing runs, we evaluated the absence of obvious batch effects in our study through principal component analysis (PCA) (Additional file 1: Fig. S11). Case and control samples were selected to approximate a balanced distribution of males and females between comparison groups to prevent sex-based methylation bias. In order to verify that case and control samples do not differ based on age, we also performed PCA using three sets of age-related CpG sites selected from literature (Methods). We showed that these CpG sites are able to differentiate PacBio samples toward different

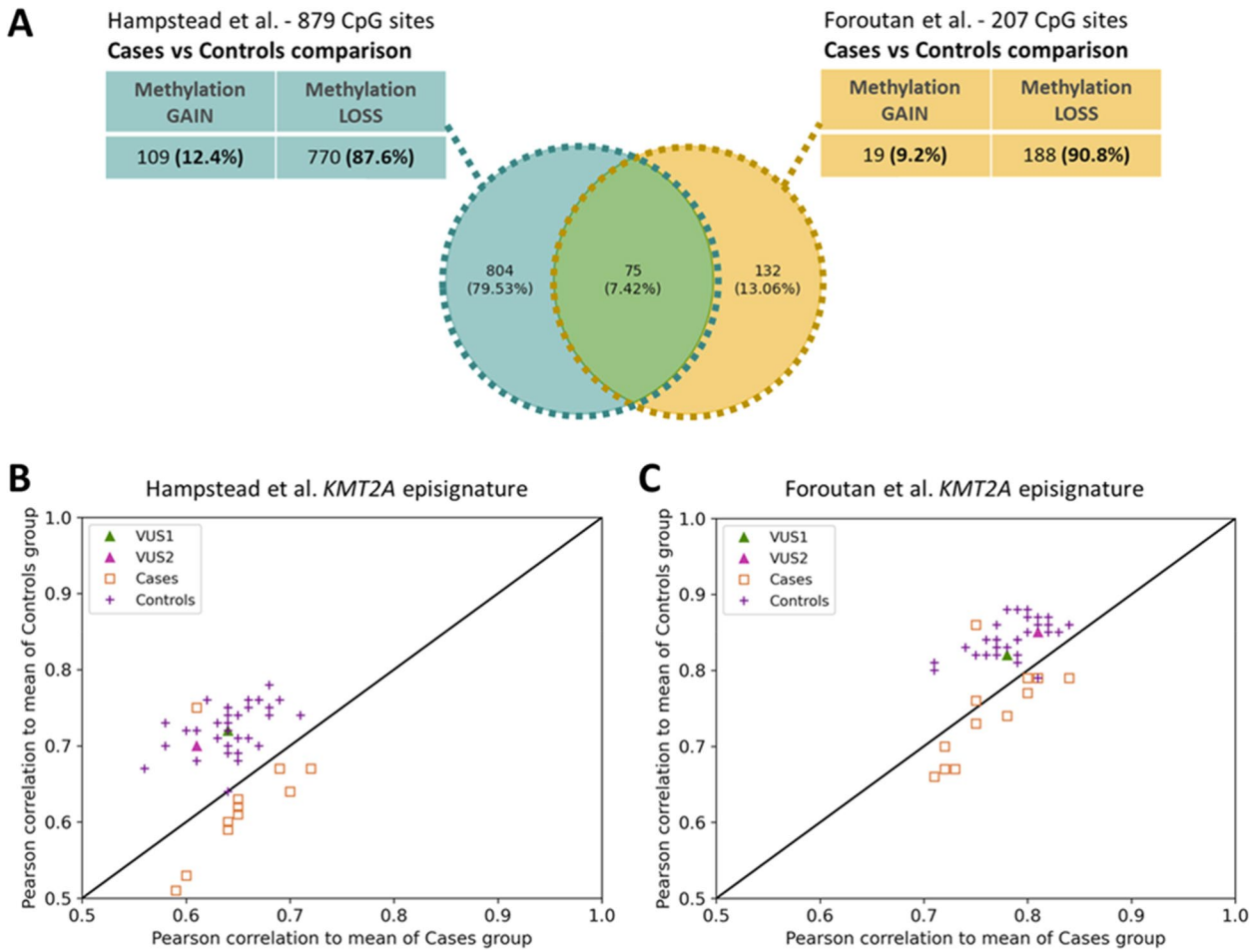

Fig. 2 Comparison of the two *KMT2A* array-based CpG signatures and classification of PacBio HiFi samples (cases/controls/VUS). **A** Description of the two sets of literature CpG sites included in the study. **B-C** Pearson correlation of methylation detection obtained for each individual PacBio HiFi sample (case, controls, VUS) compared to the mean of case and control groups using Hampstead (**B**) and Foroutan et al. (**C**) signature CpG sites

chronological age groups (Additional file 1: Fig. S12) and found that PacBio samples harboring *KMT2A* variants (cases and VUS) do not cluster separately from control samples (Additional file 1: Fig. S13).

We performed a Pearson correlation analysis between methylation detection from individual case and control samples compared to the mean of case and control groups (Fig. 2B, C). The case and control samples clustered separately, and all except one were classified within the correct group. Non-parametric classification methods (principal component analysis and hierarchical clustering) were also performed and showed overall concordant clustering of case and control samples based on the two episignatures (Additional file 1: Figs. S14, S15). A single case sample (sample 2) clustered separately from all other cases among the control group. We also applied the array-based signatures to the 2 patient samples with VUS variants in the *KMT2A* gene. For both signature sets, the two VUS samples clustered within the group of control samples (Fig. 2B, C).

Overall, these first results state toward the usability of array-based CpG episignatures on PacBio HiFi LRS data for the distinction of disease specific methylation patterns.

Episignature discovery using PacBio HiFi sequencing data

Having found that array-based *KMT2A* episignatures can distinguish cases from controls on PacBio methylation data, we further investigated whether PacBio methylation data can be used for the identification of a new *KMT2A* episignature.

PacBio HiFi methylation detection at CpG level can be used to perform differential methylation analysis

In order to assess the ability of PacBio HiFi methylation to detect episignature CpG sites, we firstly verified whether we could identify between-group methylation differences

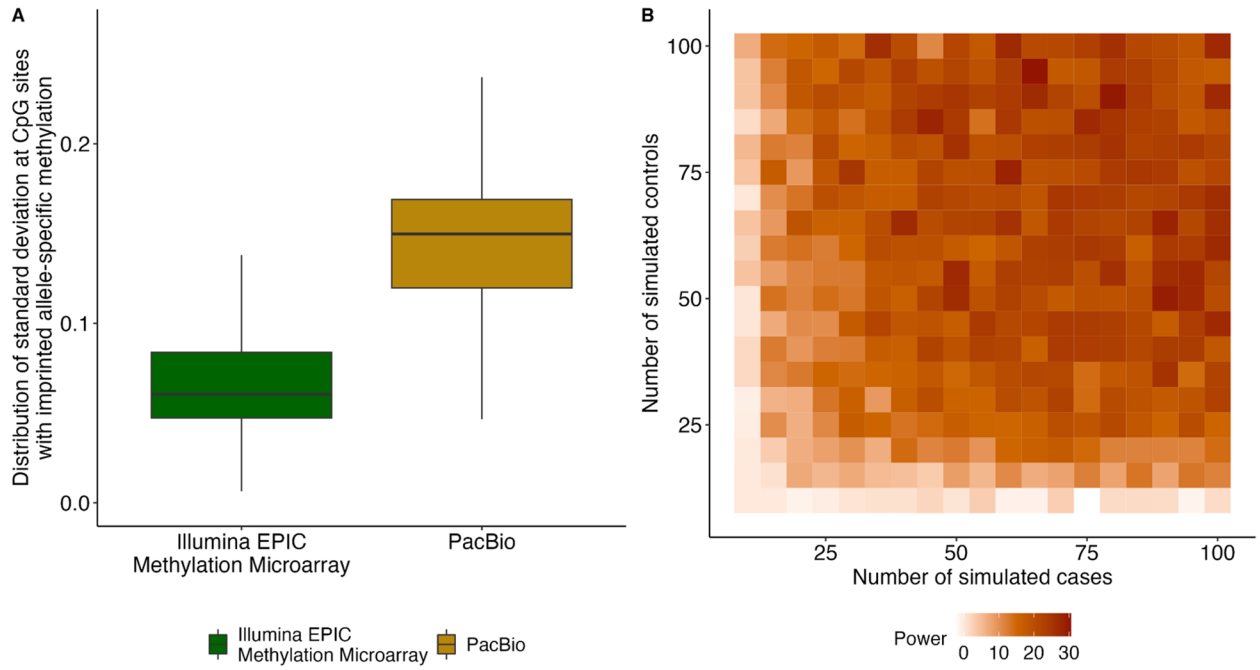

Fig. 3 Estimating power to detect robust DNA methylation signatures in PacBio long-read sequencing data. **A** Using stable imprinted sites, we assessed the standard deviation around the beta value distribution (green) and PacBio probability distribution (yellow) at CpG sites methylated on a single allele. We observe that at these sites, PacBio has a more than 2.5-fold higher standard deviation compared to the same sites profiled on methylation microarray. **B** Using sites also profiled on the Illumina EPIC array, we show that even at 100 simulated case and control samples we have only 30% power to detect CpG sites significant in both *KMT2A* array-generated DNA methylation signatures in PacBio data

based on methylation-mediated X-inactivated regions. We used 15×LRS depth of coverage sequencing data of 8 female and 29 male samples from the *KMT2A* in-house episignature cohort (Additional file 1: Table S1). The female:male ratio was set to approximate the in-house PacBio *KMT2A* case:control ratio (1:3). Initial filtering of CpG sites (6×minimal depth of coverage in all female samples and 80% of male samples) selected 19,480,361 CpG sites for analysis (Additional file 1: Table S5). Principal component analysis on genome-wide methylation detection did not show obvious genome-wide differences between the two groups of training samples prior analysis (Additional file 1: Fig. S16). The minimal difference threshold between the median methylation probabilities of female and male training groups was set to 20%.

The discovery step (Methods and Additional file 1: Table S5) identified 1500 differentially methylated CpG sites between the two groups, 1436 (95.7%) of which were located on the X chromosome. Methylation levels of these differentially methylated sites showed that they were either hypomethylated or hypermethylated in males while they presented an intermediate methylation in females (Additional file 1: Fig. S17A). The methylation signature that was based on these 1500 sites was successfully used to classify 3 independent samples consisting of 2 females (Female1_test, Female2_test) and 1 male (Male_test), which were not part of the initial training set of samples, toward the correct sex (Additional file 1: Fig. S17B, C). These results show that PacBio HiFi data can be used to detect differentially methylated sites between distinctive groups of samples. We then applied the same approach to our *KMT2A* cases and controls (Additional file 1: Tables S1, S5) while leaving one positive and negative case out. Although we were able to derive a signature, we were not able to reproduce it on the left-out samples.

Power analysis for episignature discovery from PacBio HiFi methylation data

Given our inability to detect previously identified methylation signatures in PacBio whole-genome methylation data, we hypothesized we were likely underpowered. To quantify this, we first compared the standard deviation of methylation predictions from PacBio and the Illumina EPIC methylation microarray. We used imprinted CpG sites with known allele-specific methylation (see previous) as a proxy for "heterozygous methylation." The distribution of standard deviations across imprinted CpG sites is summarized in Fig. 3A across the 529 CpG sites in imprinted regions shared by the EPIC array and PacBio samples at >6×coverage. To minimize differences between the number of samples used to calculate

the standard deviation at each CpG site, 26 PacBio controls were randomly sampled 100 times, and the lowest standard deviation for the CpG site across all samplings was used. Even so, PacBio LRS has, on average, more than two-fold higher standard deviation at imprinted CpG sites, affecting power to detect CpGs with significant methylation differences using the same methods employed on methylation microarrays for the detection of signature sites.

To estimate how many case and control samples might be required to have power to detect previously identified *KMT2A* methylation signatures in PacBio long-read sequencing data, we used only sites present in the Illumina EPIC methylation array. By restricting to this subset of probes, we can know a priori whether each CpG site is included in a previously generated *KMT2A* signature. We created case and control Gaussian probability distributions for each covered CpG site, where the mean and standard deviation at the site were derived from our *KMT2A* case and control samples respectively (see Methods). To estimate power, we used the difference between the proportion of FDR-corrected significant Mann–Whitney U p values in the 75 signature sites present in both signatures versus other sites present on the EPIC array. We show that 100 cases and 100 controls only have approximately 30% power to detect these CpG sites (Fig. 3B). Based on these calculations, more than 300 case samples would be required to have 80% power to detect *KMT2A* signatures identified on methylation microarrays, without any additional methodological improvements.

Discussion

PacBio HiFi sequencing is a technique that allows 5-methylcytosine (5mC) detection at CpG sites genome-wide. We show that PacBio can assay methylation in 5% more CpG sites genome-wide than short-read methods, including sites in complex genomic regions, centromeric sequence, ENCODE blacklist regions, and de novo CpG sites created by genetic variation. Additionally, we show that PacBio methylation predictions are concordant with short-read assays and relatively robust against sequence downsampling. We observed a linear decline in correlation between CpG calls due to downsampling, with a 0.8 correlation between 30× and 15× coverage. Using in-house PacBio trio samples, we demonstrate that PacBio can reliably detect allele-specific methylation using 25 stable imprinted DMRs, with 24 out of 25 (96%) of these regions displaying well-differentiated parent-of-origin-specific methylation patterns between haplotypes. One imprinted region (DMR ID: 16 (*SNRPN*)) lacked differential methylation for one of the two alleles across several peripheral blood samples. We showed a proper phasing of the reads on these samples but the coexistence of methylated and unmethylated reads on the same haplotypes. This observation is intriguing as it is not concordant with methylation profiles of polymorphic imprinted loci described in literature where both alleles displayed the same methylation profile, either full hypomethylation or full hypermethylation of reads [36]. Finally, we show that *KMT2A* DNA methylation episignatures generated using Illumina methylation microarrays are predictive in PacBio methylation data using 12 patients with pathogenic or likely pathogenic *KMT2A* mutations and 39 control samples.

While we were able to differentiate patients containing pathogenic *KMT2A* mutations from controls using previously published methylation microarray-derived episignatures, we were unable to generate a de novo PacBio *KMT2A* methylation signature from long-read methylation calls. To verify our approach, we identified 1500 predictive CpG sites for female versus male classification using methods conventionally employed for episignature detection on methylation microarrays. Because we expect the methylation difference between the male and female X chromosome to be between 40 and 50%, we can conclude that profiling genome-wide methylation using LRS approaches does not lead to the identification of a large number of high effect size CpG sites for the prediction of *KMT2A* variant pathogenicity not covered on methylation microarrays. In fact, we estimate that the majority of methylation differences between case and control samples in the "true" genome-wide *KMT2A* methylation signature must be considerably lower than 40%.

Interestingly, one *KMT2A* case (sample case 2) was seen to cluster discordantly toward the control group. This sample harbored a likely pathogenic *KMT2A* splice variant that was reported to the patient as the genetic cause of disease. This variant was the only variant in our cohort that was not de novo but inherited from an apparently healthy mother. Upon investigation, we concluded that this variant is unlikely to affect gene splicing and this variant was subsequently reclassified. We additionally performed LRS on the unaffected mother with the same variant and found that the mother also has no episignature resembling *KMT2A* loss-of-function. We also used array-based episignatures to classify 2 samples with missense VUS in the *KMT2A* gene. The two samples carrying *KMT2A* VUS were seen to cluster within the controls group using both signature sets, therefore making it more likely that these variants are benign rather than pathogenic according to ACMG guidelines [49].

We additionally illustrate that de novo episignature generation from PacBio long-read methylation data is challenging using existing statistical methods. CpG sites profiled using PacBio HiFi have considerably higher

standard deviations than the same sites profiled on methylation microarrays, partly due to the small number of molecules analyzed per experiment. As a result, even 100 simulated PacBio case and control samples had only about 30% power to detect CpG sites in existing *KMT2A* methylation signatures. We expect this number to be a lower bound estimate given that it was generated using only CpG sites present on methylation microarrays and using older PacBio sequencing chemistry. While sequencing additional samples can help generate predictive signatures, our findings highlight the need for new statistical methods that leverage the density of PacBio methylation data and do not rely on high-variability methylation calls at individual CpG sites. Improved methylation call accuracy, such as with ONT or PacBio's recent SPRQ chemistry upgrade, would reduce noise and enhance the reliability of the data. These advancements could provide higher-quality data needed to build robust de novo genome-wide DNA methylation episignatures for developmental disorder syndromes using much fewer than the estimated 300 case samples.

The cost of long-read sequencing is decreasing considerably. Furthermore, methylation data is generated at no additional cost for PacBio long-read genomes. Our findings suggest that PacBio sequencing may be a viable alternative to methylation microarrays as a second-tier test following whole-exome or whole-genome sequencing. Additionally, long-read whole-genome methylation profiling may also help improve the predictive ability of DNA methylation episignatures and uncover other forms of "epimutations" in rare developmental disorder syndromes not currently captured by short-read sequencing approaches.

Conclusions

We conclude that methylation calling using PacBio is highly concordant with traditional methylation assays and reliably detects imprinted regions. In addition, LRS offers advantages such as phasing information and additional methylation calls in complex genomic regions. Most importantly, we find that episignatures derived through other techniques can be readily applied on LRS methylation data to identify patients with rare diseases.

Abbreviations

5mC	5-Methylcytosine
DMRs	Differentially methylated regions
EMSeq	Enzymatic Methyl-Seq
FXS	Fragile X syndrome
GIAB	Genome In A Bottle
ICR	Imprinting control regions
IPD	Inter-pulse duration
LRS	Long-read sequencing
MethylSeq	Accel-NGS Methyl-Seq
PacBio	Pacific Biosciences
PCA	Principal component analysis
PW	Pulse width
SD	Standard deviation
VUS	Variants of uncertain significance
WSS	Wiedemann-Steiner syndrome

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-025-01506-9.

Additional file 1: Contains the supplementary figures (Supplementary Figures 1–17) and supplementary tables (Supplementary Tables 1–5).

Acknowledgements

Not applicable.

Authors' contributions

CG and JH conceived the study. RD, AO, GK, SH, RT, JCG, RP, TH, and HY generated the data described in this manuscript. VI, MG, CG, and JH analyzed and interpreted the data in this study. VI, MG, LV, AH, CG, and JH wrote and edited the manuscript. All authors read and approved the final manuscript.

Funding

Not applicable.

Data availability

Data is provided within the manuscript or supplementary information files. Methylation probabilities at GRCh38 reference CpG sites (excluding those at SNPs >1% MAF in gnomAD v4 ancestral populations) for all samples are available upon request.) accession ID EGAS00001008250 [51] Access is controlled and available upon request and approval by the data access committee. All code used to process data and generate files is publicly available on Github at https://github.com/Genome-Bioinformatics-RadboudUMC/Pacbio-Methylation-Episignatures [50] under GPL-3.0 license.

Declarations

Ethics approval and consent to participate

Diagnostic laboratories can use (de-identified) samples from archived clinical samples to validate and optimize diagnostic assays as approved by the institutional review board "Commissie Mensgebonden Onderzoek Regio Arnhem-Nijmegen" under number (2020-7142). Therefore, the individuals in this study are shown without information about age or other identifiers besides biological sex. These methods are also in accordance with relevant guidelines and regulations of the Declaration of Helsinki.

Consent for publication

Patient samples, together with a basic phenotype description, were anonymized. The study was approved by the institutional review board "Commissie Mensgebonden Onderzoek Regio Arnhem-Nijmegen" under number 2011/188. All patients in this study have consented to their anonymized data being made available for research use.

Competing interests

Part of the reagents for this study were provided by Pacific Biosciences. The authors declare that they have no competing interests

Author details

[1]Department of Human Genetics, Radboud University Medical Center, Geert Grooteplein Zuid 10, Nijmegen, GA 6525, The Netherlands. [2]Département de Génétique Médicale, Institut Fédératif de Biologie, Hôpital Purpan, Toulouse 31000, France. [3]Radboudumc Research Institute for Medical Innovation, Radboud University Medical Center, Nijmegen, The Netherlands. [4]Department of Laboratory Medicine, Translational Metabolic Laboratory, Radboud University Medical Center, Nijmegen, The Netherlands.

Received: 16 September 2024 Accepted: 29 June 2025
Published online: 26 January 2026

References

1. Greenberg MVC, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. Nat Rev Mol Cell Biol. 2019;20(10):590–607. https://doi.org/10.1038/s41580-019-0159-6.
2. Lam LL, Emberly E, Fraser HB, et al. Factors underlying variable DNA methylation in a human community cohort. Proc Natl Acad Sci. 2012;109(supplement_2):17253–60. https://doi.org/10.1073/pnas.1121249109.
3. Monk D, Mackay DJG, Eggermann T, Maher ER, Riccio A. Genomic imprinting disorders: lessons on how genome, epigenome and environment interact. Nat Rev Genet. 2019;20(4):235–48. https://doi.org/10.1038/s41576-018-0092-0.
4. Eggermann T, Perez de Nanclares G, Maher ER, et al. Imprinting disorders: a group of congenital disorders with overlapping patterns of molecular changes affecting imprinted loci. Clin Epigenetics. 2015;7(1):123. https://doi.org/10.1186/s13148-015-0143-8.
5. Cerrato F, Sparago A, Ariani F, et al. DNA methylation in the diagnosis of monogenic diseases. Genes (Basel). 2020;11(4):355. https://doi.org/10.3390/genes11040355.
6. Martinez-Delgado B, Barrero MJ. Epigenomic approaches for the diagnosis of rare diseases. Epigenomes. 2022;6(3):21. https://doi.org/10.3390/epigenomes6030021.
7. Aref-Eshghi E, Bend EG, Colaiacovo S, et al. Diagnostic utility of genome-wide DNA methylation testing in genetically unsolved individuals with suspected hereditary conditions. Am J Hum Genet. 2019;104(4):685–700. https://doi.org/10.1016/j.ajhg.2019.03.008.
8. 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. https://doi.org/10.3390/ijms21239303.
9. Gouil Q, Keniry A. Latest techniques to study DNA methylation. Essays Biochem. 2019;63(6):639–48. https://doi.org/10.1042/EBC20190027.
10. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 2012;13(1):36–46. https://doi.org/10.1038/nrg3117.
11. 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(8):1359–66. https://doi.org/10.1016/j.ajhg.2021.06.015.
12. Pidsley R, Zotenko E, Peters TJ, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biol. 2016;17(1):208. https://doi.org/10.1186/s13059-016-1066-1.
13. Sun Z, Cunningham J, Slager S, Kocher JP. Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis. Epigenomics. 2015;7(5):813–28. https://doi.org/10.2217/epi.15.21.
14. Flusberg BA, Webster DR, Lee JH, et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods. 2010;7(6):461–5. https://doi.org/10.1038/nmeth.1459.
15. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. Detecting DNA cytosine methylation using nanopore sequencing. Nat Methods. 2017;14(4):407–10. https://doi.org/10.1038/nmeth.4184.
16. Ni P, Nie F, Zhong Z, et al. DNA 5-methylcytosine detection and methylation phasing using PacBio circular consensus sequencing. Nat Commun. 2023;14(1):4054. https://doi.org/10.1038/s41467-023-39784-9.
17. Silva C, Machado M, Ferrão J, Sebastião Rodrigues A, Vieira L. Whole human genome 5′-mC methylation analysis using long read nanopore sequencing. Epigenetics. 2022;17(13):1961–75. https://doi.org/10.1080/15592294.2022.2097473.
18. Sigurpalsdottir BD, Stefansson OA, Holley G, et al. A comparison of methods for detecting DNA methylation from long-read sequencing of human genomes. Genome Biol. 2024;25(1):69. https://doi.org/10.1186/s13059-024-03207-9.
19. Foox J, Nordlund J, Lalancette C, et al. The SEQC2 epigenomics quality control (EpiQC) study. Genome Biol. 2021;22(1):332. https://doi.org/10.1186/s13059-021-02529-2.
20. Zook JM, Catoe D, McDaniel J, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data. 2016;3(1): 160025. https://doi.org/10.1038/sdata.2016.25.
21. PacificBiosciences. HG002 public repository. 2022. https://downloads.pacbcloud.com/public/dataset/HG002-CpG-methylation-202202.
22. PacificBiosciences. CCS: generate highly accurate single-molecule consensus reads (HiFi reads). Github. 2018. https://github.com/PacificBiosciences/ccs.
23. PacificBiosciences. Primrose. Github. 2023. https://github.com/PacificBiosciences/primrose.
24. PacificBiosciences. pbmm2: a minimap2 frontend for PacBio native data formats. Github. 2022. https://github.com/PacificBiosciences/pbmm2.
25. Poplin R, Chang PC, Alexander D, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36(10):983–7. https://doi.org/10.1038/nbt.4235.
26. Patterson M, Marschall T, Pisanti N, et al. WhatsHap: weighted haplotype assembly for future-generation sequencing reads. J Comput Biol. 2015;22(6):498–509. https://doi.org/10.1089/cmb.2014.0157.
27. PacificBiosciences. pb-cpg-tools. Github. 2023. https://github.com/PacificBiosciences/pb-cpg-tools.
28. Foox, J. et al. EpiQC. Gene Expression Omnibus. 2021. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE186383.
29. Pedersen BS, Eyring K, De S, Yang I V., Schwartz DA. Fast and accurate alignment of long bisulfite-seq reads. Published online 6 Jan 2014.
30. Ryan, D. MethylDackel: a (mostly) universal methylation extractor for BS-seq experiments. Github. 2016. https://github.com/dpryan79/MethylDackel.
31. PacificBiosciences. pb-human-wgs-workflow-snakemake: workflow for the comprehensive detection and prioritization of variants in human genomes with PacBio HiFi reads. Github. 2023. https://github.com/PacificBiosciences/pb-human-wgs-workflow-snakemake.
32. Wenger AM, Peluso P, Rowell WJ, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019;37(10):1155–62. https://doi.org/10.1038/s41587-019-0217-9.
33. 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). https://doi.org/10.1093/bioinformatics/btae042.
34. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2). https://doi.org/10.1093/gigascience/giab008.
35. Perez G, Barber GP, Benet-Pages A, et al. The UCSC Genome Browser database: 2025 update. Nucleic Acids Res. 2025;53(D1):D1243–9. https://doi.org/10.1093/nar/gkae974.
36. Joshi RS, Garg P, Zaitlen 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(3):555–66. https://doi.org/10.1016/j.ajhg.2016.06.032.
37. Jima DD, Skaar DA, Planchart A, et al. Genomic map of candidate human imprint control regions: the imprintome. Epigenetics. 2022;17(13):1920–43. https://doi.org/10.1080/15592294.2022.2091815.
38. Salameh Y, Bejaoui Y, El Hajj N. DNA methylation biomarkers in aging and age-related diseases. Front Genet. 2020;11. https://doi.org/10.3389/fgene.2020.00171.
39. Hannum G, Guinney J, Zhao L, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013;49(2):359–67. https://doi.org/10.1016/j.molcel.2012.10.016.
40. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14(10):3156. https://doi.org/10.1186/gb-2013-14-10-r115.
41. Levine ME, Lu AT, Quach A, et al. An epigenetic biomarker of aging for lifespan and healthspan. Aging (Albany NY). 2018;10(4):573–91. https://doi.org/10.18632/aging.101414.
42. Foroutan A, Haghshenas S, Bhai P, et al. Clinical utility of a unique genome-wide DNA methylation signature for KMT2A-related syndrome. Int J Mol Sci. 2022;23(3):1815. https://doi.org/10.3390/ijms23031815.
43. Hampstead J. Functional genomics of developmental disorders. Univ Cambridge Apollo-University Cambridge Repos. Published online 2023. https://doi.org/10.17863/CAM.107502.

44. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17(3):261–72. https://doi.org/10.1038/s41592-019-0686-2.
45. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24–6. https://doi.org/10.1038/nbt.1754.
46. Kucuk E, van der Sanden BPGH, O'Gorman L, et al. Comprehensive de novo mutation discovery with HiFi long-read sequencing. Genome Med. 2023;15(1):34. https://doi.org/10.1186/s13073-023-01183-6.
47. Levy MA, McConkey H, Kerkhof J, et al. Novel diagnostic DNA methylation episignatures expand and refine the epigenetic landscapes of Mendelian disorders. Hum Genet Genomics Adv. 2022;3(1): 100075. https://doi.org/10.1016/j.xhgg.2021.100075.
48. Sheppard SE, Campbell IM, Harr MH, et al. Expanding the genotypic and phenotypic spectrum in a diverse cohort of 104 individuals with Wiedemann-Steiner syndrome. Am J Med Genet Part A. 2021;185(6):1649–65. https://doi.org/10.1002/ajmg.a.62124.
49. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405–24. https://doi.org/10.1038/gim.2015.30.
50. Ivashchenko V, et al. Genome-wide methylation detection and episignature analysis using PacBio long-read sequencing. Github. 2025. https://github.com/Genome-Bioinformatics-RadboudUMC/Pacbio-Methylation-Episignatures.
51. Genome-wide methylation detection and episignature analysis using PacBio long-read sequencing. https://ega-archive.org/studies/EGAS00001008250.

Publisher's Note