Genome-wide CpG methylation calling with standard HiFi whole genome sequencing

Justin Blethrow, Richard Hall, Kristofor Nyquist, Aaron Wenger, Daniel Portik
PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

Introduction

PacBio HiFi sequencing provides the most accurate and complete characterization of human genomes. Sequencing observes a polymerase in real time as it incorporates fluorescently labeled nucleotides to synthesize a DNA strand. Kinetic signatures including pulse width and interpulse duration correlate with chemical modifications to the canonical DNA bases (Fig. 1), including the 5-methylcytosine (5mC) modification without bisulfite treatment.

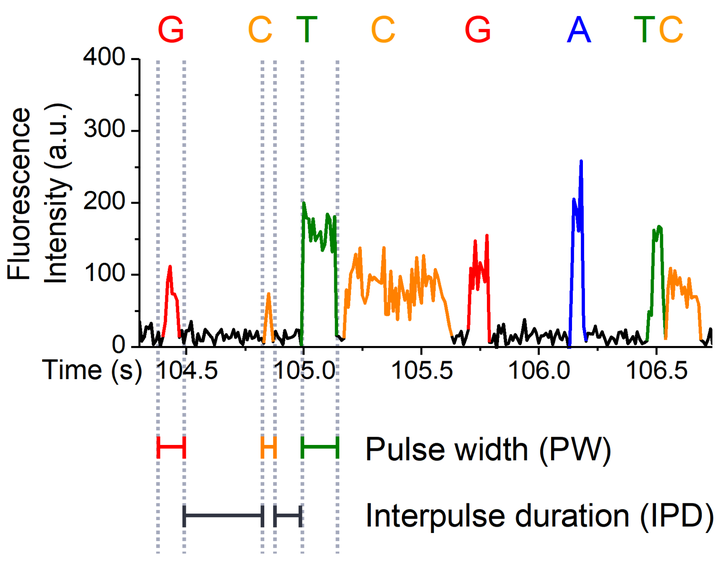

Figure 1. Kinetic signatures. Example trace showing pulse width (time of incorporation) and interpulse duration (time between adjacent incorporations). Image modified from Flusberg et al. (2010)[1].

Methods

HiFi sequencing observes the same molecule across multiple serial passes (Fig. 2), opening new approaches to detect 5mC. We implemented a multilayer convolution neural network to combine kinetics from multiple passes and assign a probability of methylation to each CpG. We trained the model on fully unmethylated (whole-genome amplification) and fully methylated (M.SssI-treated) reads. The training uses all sequence contexts from the reads, but does not require the reads to be aligned to a reference genome.

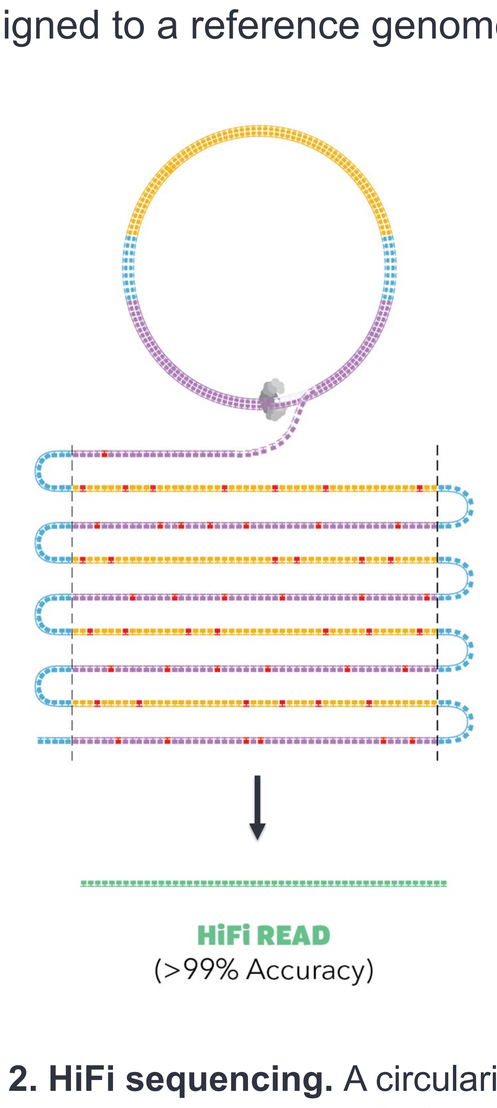

Figure 2. HiFi sequencing. A circularized template is sequenced with multiple passes. The subreads are used to produce highly accurate consensus sequence, or HiFi read, with 99.9% accuracy (QV 30). No library modification is required to obtain the kinetics information to predict 5mC.

Workflow

A model implemented in the **primrose** software predicts 5mC probabilities for HiFi reads (Fig. 3). The SAM tags encoding 5mC positions and scores (MM, ML) are added to all HiFi reads.

Feature vector

16	Forward IPD
16	Forward PW
16	Reverse IPD
16	Reverse PW
1	Passes

Convolutional neural network

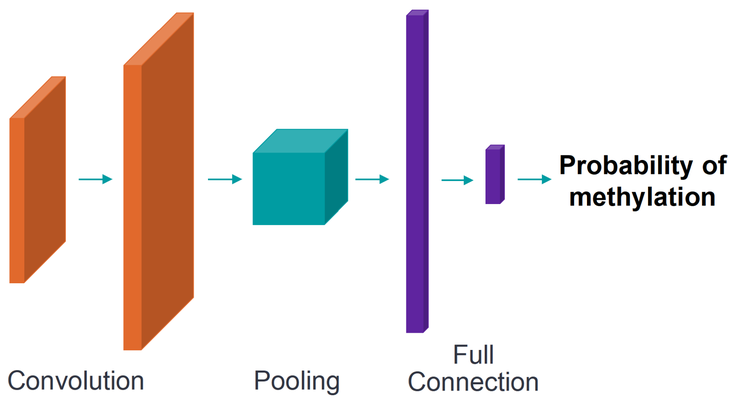

Figure 3. Primrose overview. Visualization of the feature vector and neural network implemented in Primrose.

The HiFi reads with 5mC tags (supplied in an unaligned BAM format) can be aligned to a reference using **pbmm2.** From the alignments, pileup scores for 5mC across CpG sites can be obtained using PacBio's **CpG tools**. If reads are phased, 5mC pileup scores are also provided for each haplotype.

The above tools are available on **github**.

primrose
- PacificBiosciences/primrose

pbmm2
- PacificBiosciences/pbmm2

CpG tools
- PacificBiosciences/pb-CpG-tools

Validation

We sequenced multiple Genome in a Bottle (GIAB) samples and performed the 5mC workflow. The HiFi CpG methylation calls have a high correlation with calls from orthogonal technologies[2], including EMSeq, MethylSeq, and ONT (Fig. 4).

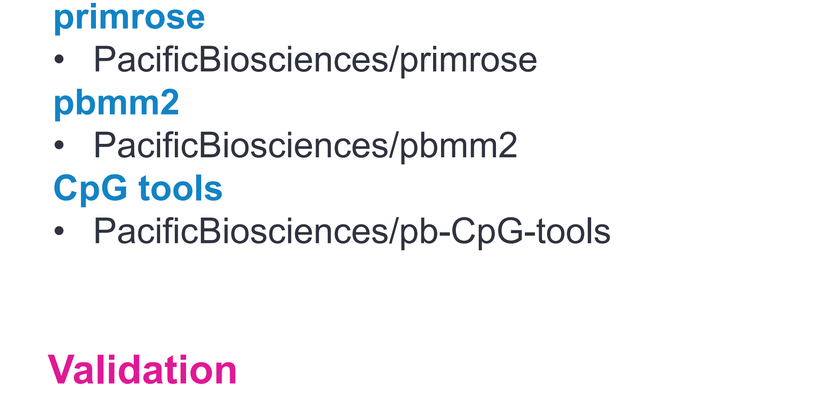

Figure 4. Technology comparison. Pearson correlation by position, compared across technologies and GIAB samples. HiFi datasets were ~30x depth of coverage per sample.

Demonstrations

We used HiFi reads with 5mC to detect:
- hypermethylation associated with a pathogenic repeat expansion (Fig. 5)
- parental imprinting revealed by haplotype phasing (Fig. 6)
- uniparental heterodisomy (Fig. 7)

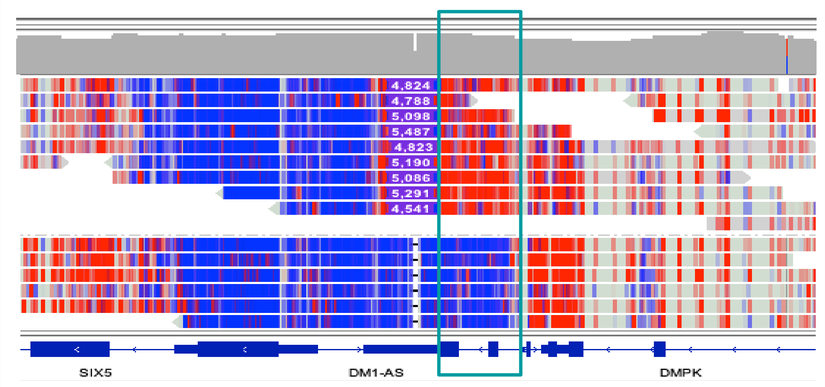

Figure 5. Repeat expansions in *DMPK*. Myotonic dystrophy due to 4.5–5.5 kb repeat expansions which induced hypermethylation. Region shown is ~8.5 kb. Example courtesy of Tomi Pastinen, Children's Mercy Kansas City.

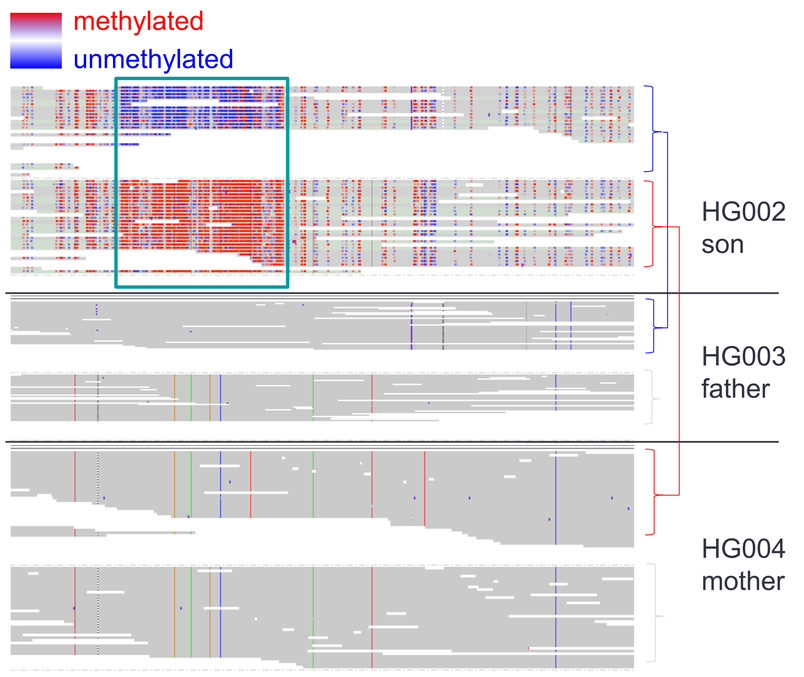

Figure 6. Parental imprinting in *PEG3*. A 15 kb region of *PEG3* is shown for a trio of samples. The HiFi reads allow haplotype phasing and correct identification of the maternal and paternal alleles. The paternally inherited allele in HG002 is clearly shown to be hypomethylated.

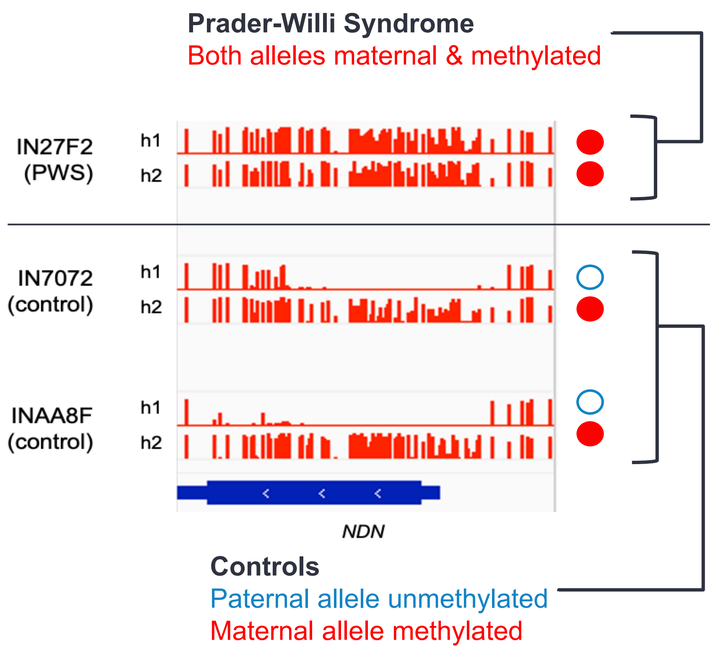

Figure 7. Uniparental heterodisomy. Prader-Willi Syndrome due to presence of two maternal alleles which display hypermethylation. A 1 kb window containing *NDN* on chr15 is shown for two control samples and the affected individual. Example courtesy of Matthew Bainbridge, Rady Children's Institute of Genomic Medicine.

Conclusions

We demonstrate the ability to accurately detect 5mC in CpG with HiFi sequencing of samples prepared using standard libraries without bisulfite treatment.

References

1. Flusberg B, et al. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods*, 7(6), 461–465.

2. Foox J, et al. (2021). The SEQC2 epigenomics quality control (EpiQC) study. *Genome Biology,* 22, 332.