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.

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.

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.

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.

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, including EMSeq, MethylSeq, and ONT (Fig. 4).

Conclusions

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

References


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 1. Kinetic signatures. Example trace showing pulse width (time of incorporation) and interpulse duration without bisulfite treatment.

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

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

Figure 5. Repeat expansions in DMPK. Myotonic dystrophy due to 4.5–5.5 kb repeat expansions which induced hypermethylation. Region shown is ~3.5 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 DMPK is shown for two control samples and the affected individual. Example courtesy of Matthew Bainbridge, Rady Children’s Institute of Genomic Medicine.