PacBio HiFi sequencing provides highly accurate CpG methylation calls without bisulfite treatment

Christopher T Saunders, Daniel Portik, Richard J Hall, David Seifert, Darien Diaz, Armin Töpfer, Kristofor Nyquist, Aaron M Wenger
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. This kinetic information is independent of the fluorescence intensity used for base calling.

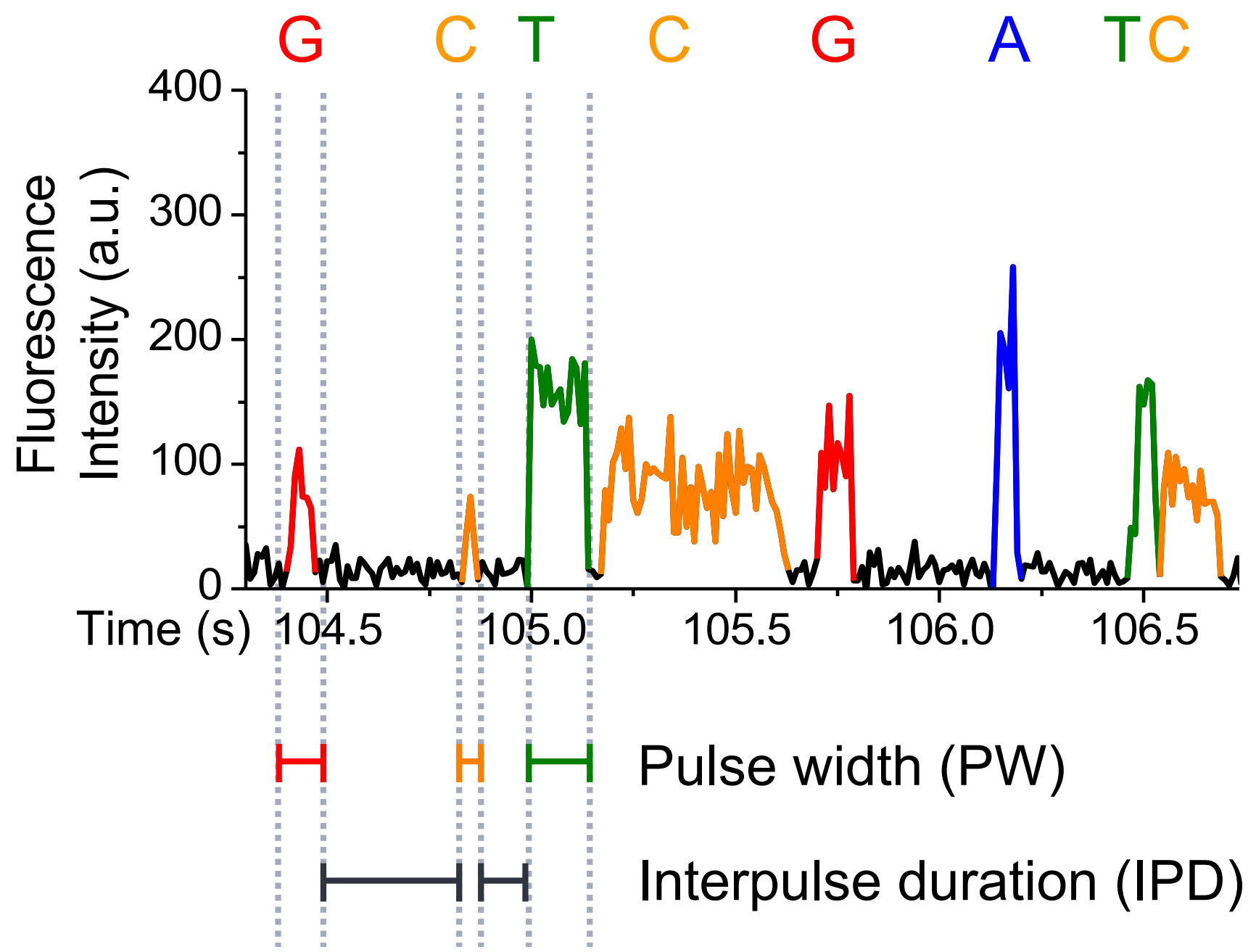


Figure 1. Kinetic signatures. Example trace showing pulse width (time of incorporation) and interpulse duration (time between adjacent incorporations). Image modified from Flusberg et al. (2010).[1]

Methods

HiFi sequencing observes the same molecule across multiple serial passes (Fig. 2), opening new approaches to detect 5mC.

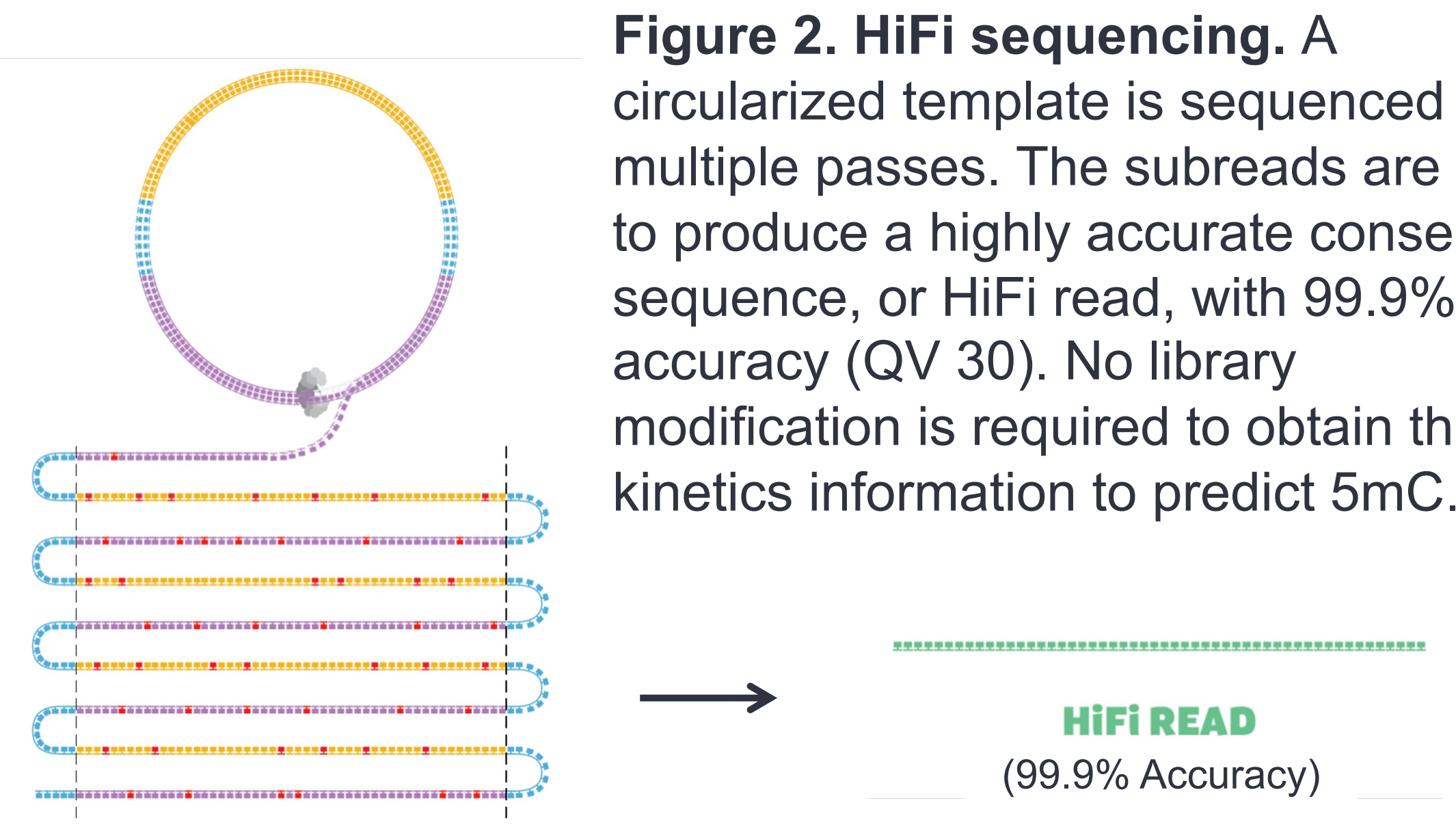


Figure 2. HiFi sequencing. A circularized template is sequenced with multiple passes. The subreads are used to produce a highly accurate consensus sequence, or HiFi read, with 99.9% accuracy (QV 30). No library modification is required to obtain the kinetics information to predict 5mC.

We implemented a multilayer convolutional neural network to combine kinetics from multiple passes and assign a methylation probability to each CpG (Fig. 3). 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.

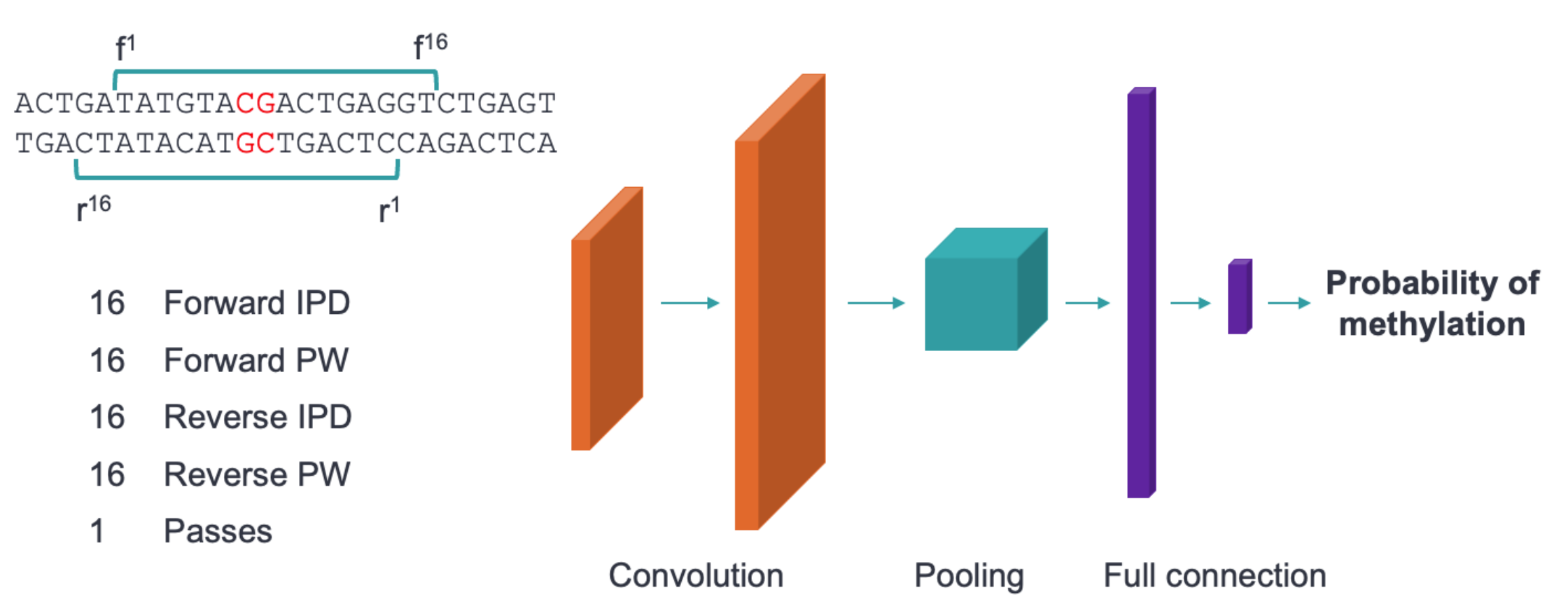
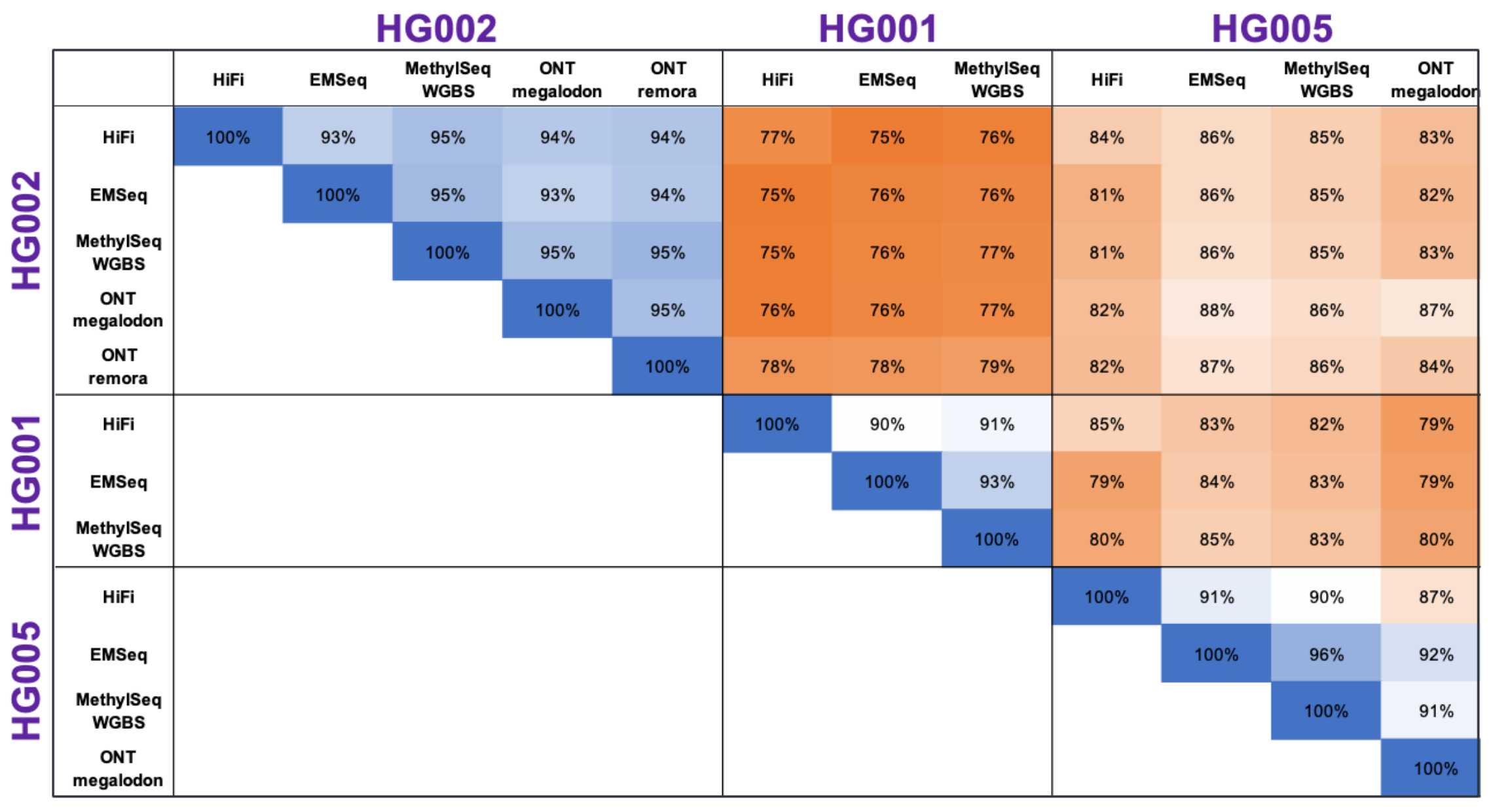


Figure 3. Methylation model overview. Visualization of the feature vector and neural network architecture.

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] (Fig. 4). HiFi reads do not exhibit depth bias against CpG islands, in contrast to depth observations from WGBS (Fig. 5). HiFi methylation signals also recapitulate known biological phenomena, such as parental imprinting in *PEG3* (Fig. 6).



		HG002					HG001			HG005			
		HiFi	EMSeq	MethylSeq WGBS	ONT megalodon	ONT remora	HiFi	EMSeq	MethylSeq WGBS	HiFi	EMSeq	MethylSeq WGBS	ONT megalodon
HG002	HiFi	100%	93%	95%	94%	94%	77%	75%	76%	84%	86%	85%	83%
	EMSeq		100%	95%	93%	94%	75%	76%	76%	81%	86%	85%	82%
	MethylSeq WGBS			100%	95%	95%	75%	76%	77%	81%	86%	85%	83%
	ONT megalodon				100%	95%	76%	76%	77%	82%	88%	86%	87%
	ONT remora					100%	78%	78%	79%	82%	87%	86%	84%
HG001	HiFi						100%	90%	91%	85%	83%	82%	79%
	EMSeq							100%	93%	79%	84%	83%	79%
	MethylSeq WGBS								100%	80%	85%	83%	80%
HG005	HiFi									100%	91%	90%	87%
	EMSeq										100%	96%	92%
	MethylSeq WGBS											100%	91%
	ONT megalodon												100%

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

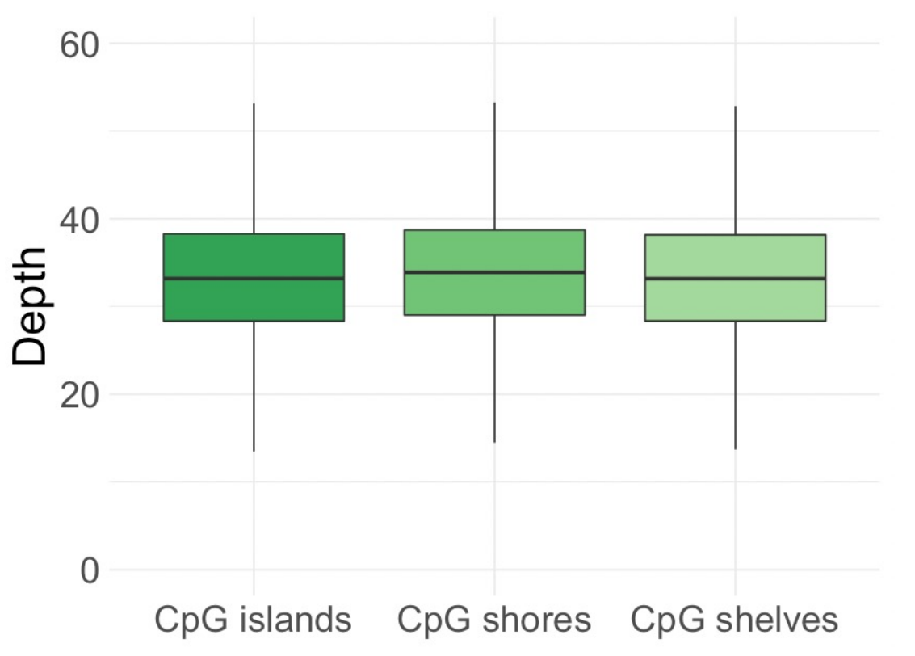


Figure 5. Coverage distribution. HiFi read depth in CpG islands, shores, and shelves is shown for HG002. HiFi reads do not exhibit depth bias against CpG islands, as observed with WGBS.[2]

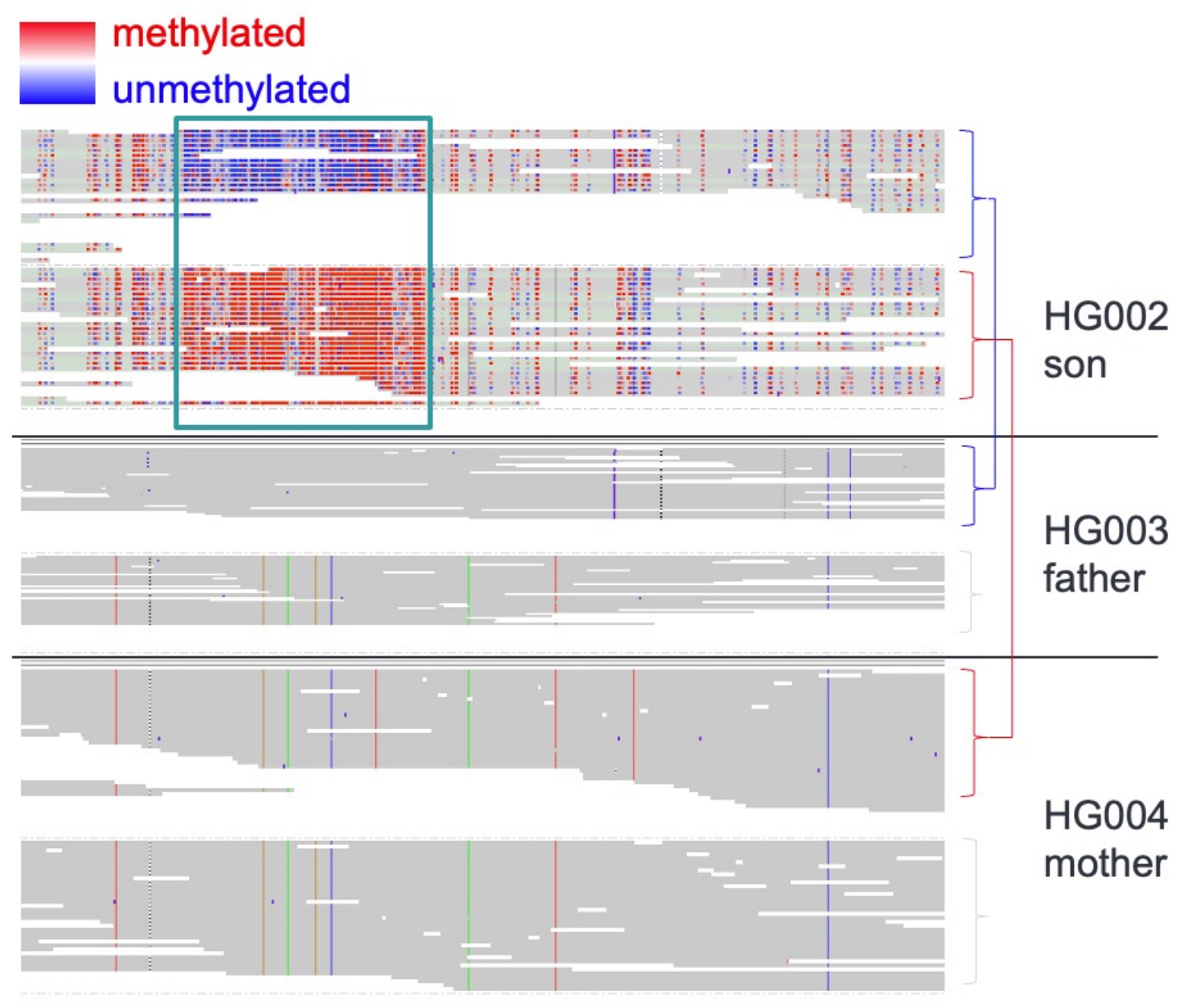


Figure 6. Parental imprinting in *PEG3*. A 15 kb region of *PEG3* is shown for a trio of samples. The HiFi reads allow haplotype phasing and identification of the maternal and paternal alleles. The paternally inherited allele in HG002 is clearly shown to be hypomethylated.

Available software

A model implemented in the **primrose** software predicts 5mC probabilities for HiFi reads. 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.

The above tools are available at **https://github.com/PacificBiosciences**.

primrose: PacificBiosciences/primrose
pbmm2: PacificBiosciences/pbmm2
CpG tools: PacificBiosciences/pb-CpG-tools

Demonstrations

HiFi reads with 5mC are used to detect:

- Hypermethylation associated with a pathogenic repeat expansion (Fig. 7)
- Imprinting disorder associated with Prader-Willi syndrome (Fig. 8)

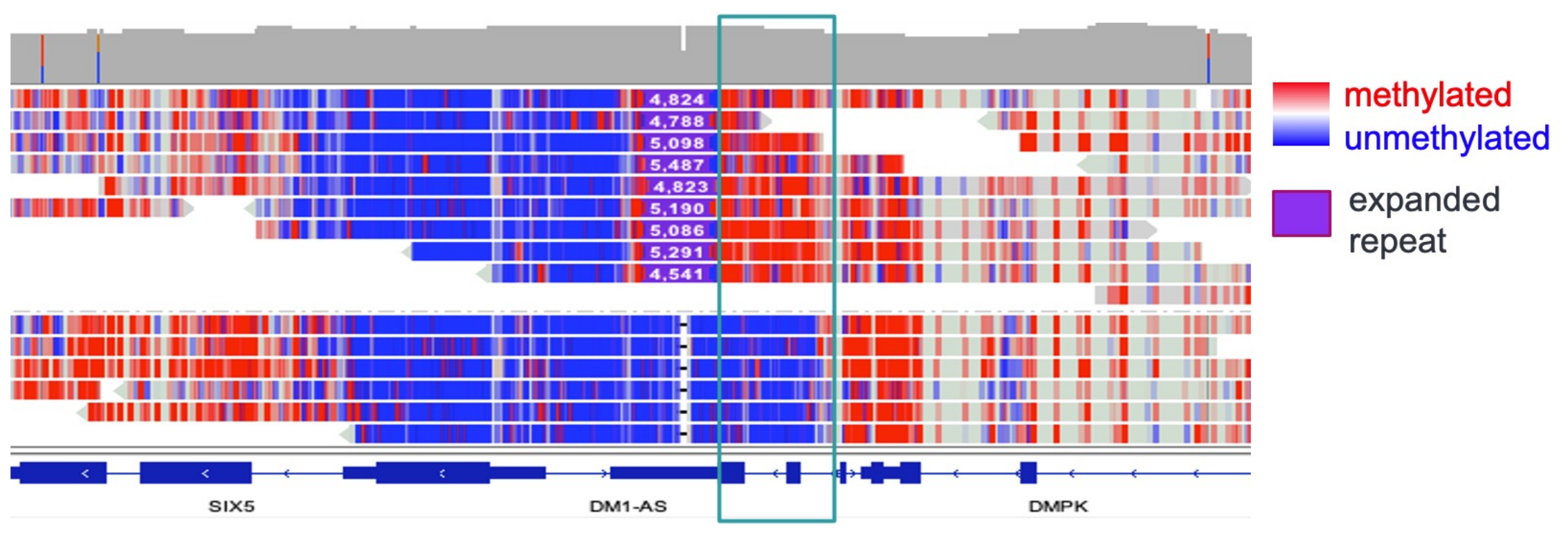


Figure 7. Repeat expansions in *DMPK*. Myotonic dystrophy due to 4.5–5.5 kb repeat expansions which induced hypermethylation. Region shown is ~8.5 kb. Example courtesy of Tomi Pastinen, Children’s Mercy Kansas City.

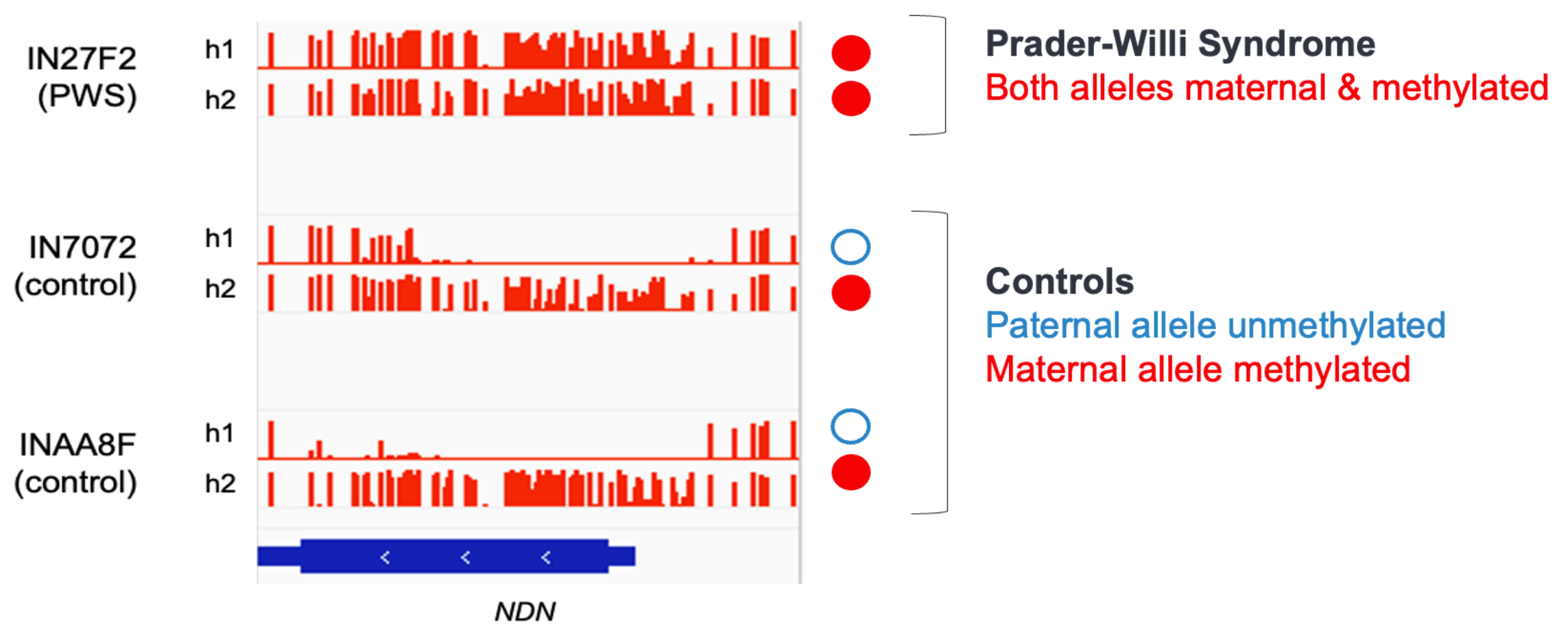


Figure 8. Uniparental heterodisomy. Prader-Willi syndrome due to presence of two maternal alleles which display hypermethylation. A 1 kb window containing *NDN* on chr15 is shown for two control samples and the affected individual. Example courtesy of Matthew Bainbridge, Rady Children’s Institute of Genomic Medicine.

Future capabilities

New base modification capabilities are under development:

- 5mC accuracy improvements and progression towards generalized calling in all contexts (CpG/CHG/CHH) (Fig. 9)
- 6mA calling for HiFi reads to support single-molecule footprinting and microbial assays (Fig. 10)

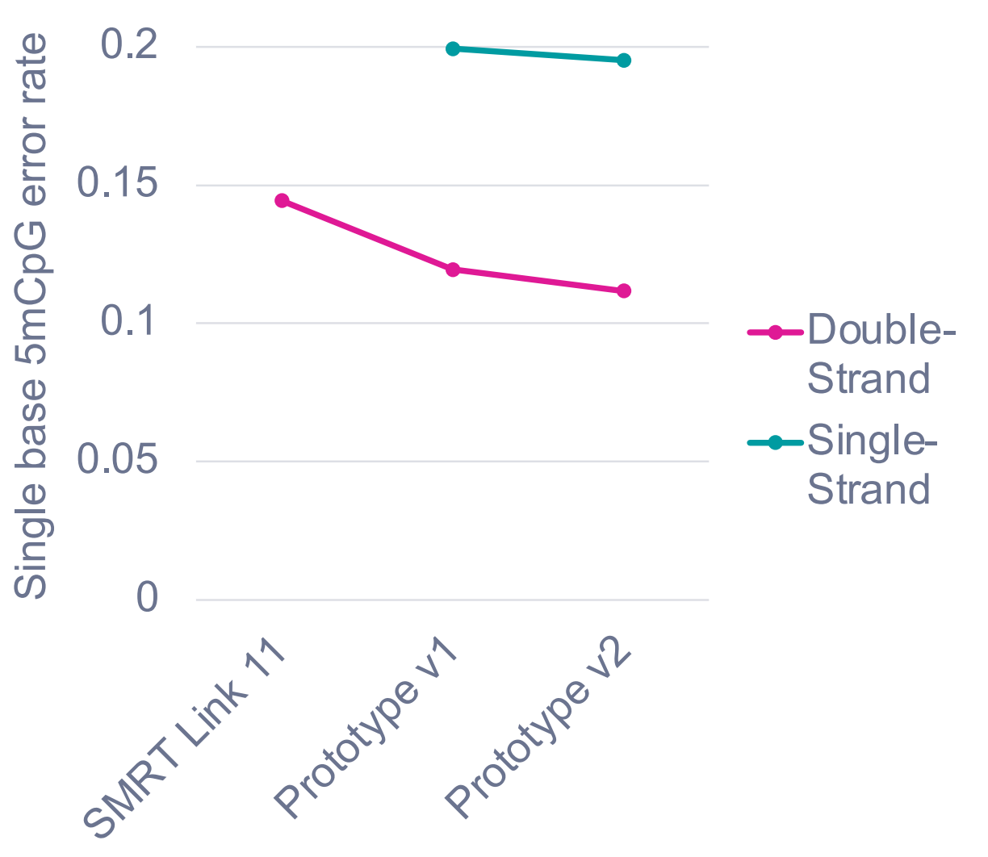


Figure 9. New 5mC prototype model accuracy. New 5mC model prototypes extend current CpG caller (SMRT Link 11) with higher accuracy double-strand calling and new single-strand capability as a first step towards generalized calling beyond CpG contexts.

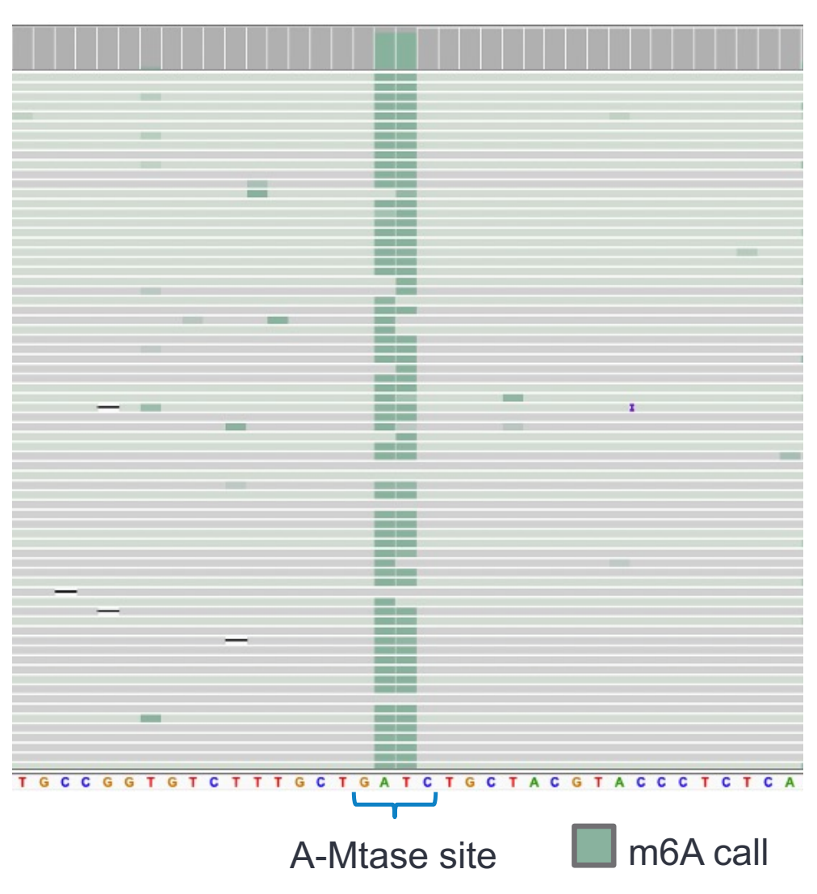


Figure 10. 6mA calling for HiFi reads. HiFi reads with 6mA annotations are shown for a segment of the *E. coli* genome centered on a *Dam* methyltransferase site (GATC). New prototype model directly annotates 6mA probabilities in HiFi BAM files with kinetic data.

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.