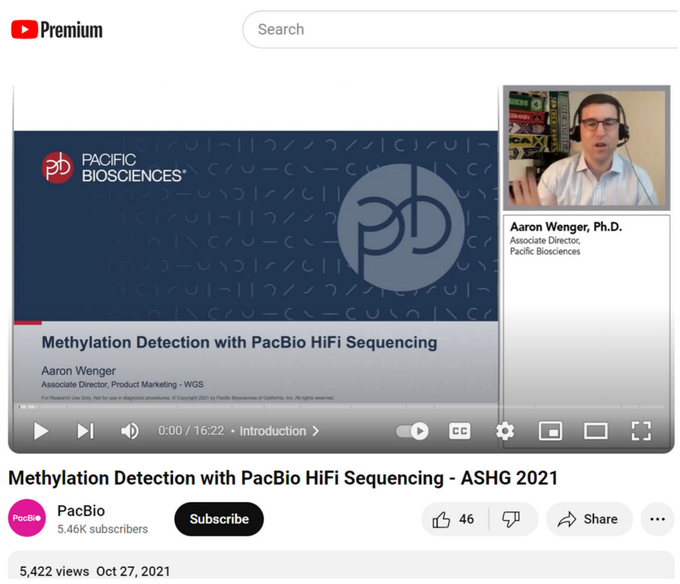
EXHIBIT 2054
Comparison of '794 Patent Claim 1, EX1030, EX2024, and EX2047.

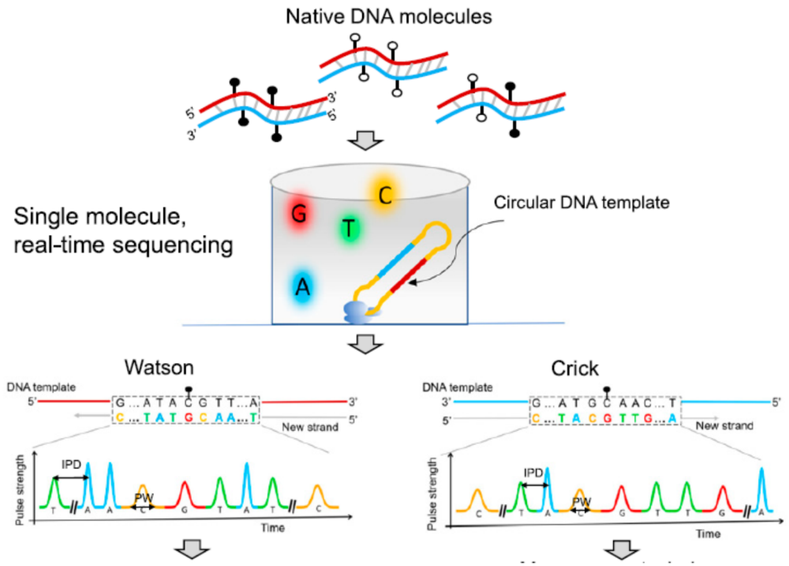
'794 Patent Claim 1	**Tse EX1030**	**Dr. Wenger's 2021 ASHG Annual Meeting Presentation, "Methylation Detection with PacBio HiFi Sequencing" (EX2024)**	**PacBio's Primrose Poster (EX2047)**
1. A method for detecting a modification of a nucleotide in a nucleic acid molecule, the method comprising:	EX1030 at 1, Title: "Genome-wide detection of cytosine methylation by single molecule real-time sequencing." EX1030 at 1, Abstract: "Using single molecule real-time sequencing, we developed a methodology to directly examine 5mC."	Presentation title: "Methylation Detection with PacBio HiFi Sequencing," Title Slide, EX2024 at 0:00-0:10.  EX2024 screenshot at 0:00.	Title: "Genome-wide CpG methylation calling with standard HiFi whole genome sequencing," EX2047.

(a) receiving data acquired by measuring pulses in an optical signal corresponding to nucleotides sequenced in a sample nucleic acid molecule and obtaining, from the data, values for the following properties:

for each nucleotide:

an identity of the nucleotide, a position of the nucleotide within the sample nucleic acid molecule, a width of the pulse corresponding to the nucleotide, and

an interpulse duration representing a time between the pulse corresponding to the nucleotide and a pulse corresponding to a neighboring nucleotide;

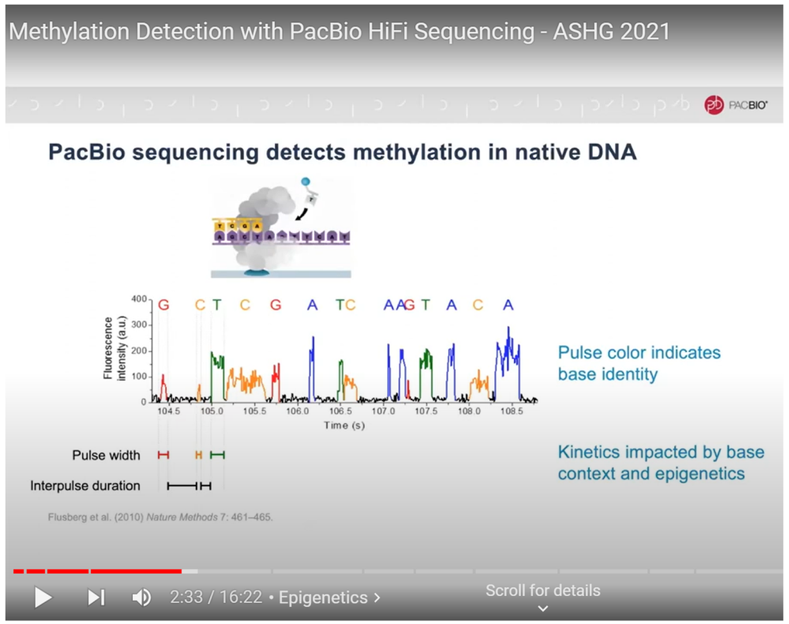


EX1030 at 2, Fig. 1 (partial)

"Double-stranded DNA molecules were ligated with hairpin adapters, forming circular DNA templates. DNA polymerase in a ZMW would incorporate nucleotides labeled with different fluorophores into the complementary strand of a DNA template, thus emitting different fluorescent colors indicating nucleotide information: for example, red, yellow, green, and blue colors represented G, C, T, and A, respectively. The light pulse signals were reflective of DNA polymerase kinetics, depending on the base modifications. Pulse signals included IPD and PW." EX1030 at 2, Fig. 1 legend.

"Different fluorescent dyes were used to determine the base content. For example, red, yellow, green, and blue colors represented G, C, T, and A, respectively (Fig. 1). The light pulse signals emitted from fluorescently labeled nucleotides were reflective of DNA polymerase kinetics, depending on the base modifications. Thus, the appropriate use of pulse signals would make it possible to determine whether a cytosine was methylated or not. Pulse signals included the IPD, that represented the time duration between two consecutive base incorporations, and the pulse width (PW), that represented the time duration of the emission of fluorescent signal associated with a base incorporation. The pulse signals were associated with the sequence context in which the polymerization reaction was occurring. Herein, we developed an approach for determining DNA methylation by using pulse signals, including IPDs, PWs and the sequence context. Sequence context referred to the base compositions (A, C, G, or T) and the base orders in a stretch of DNA." EX1030 at 3.

Dr. Wenger described that PacBio's new tool under development to call 5mC would receive the HiFi read data from PacBio sequencing of "native DNA," which were optical signals in color and were used to obtain nucleotide identity and position, IPD, and PW of the native DNA. EX2024 at 2:27-3:11.



EX2024 screenshot at 2:33.

Identity of nucleotides: "Pulse color indicates base identity." EX2024 screenshot at 2:33.

"Position of the nucleotide within the sample nucleic acid molecule," IPD and PW: PacBio shows that it captures the positions of the bases over the time of the sequencing read, as well as measurement of IPD and PW. EX2024 screenshot at 2:33.

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.

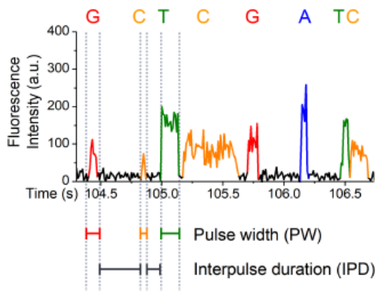
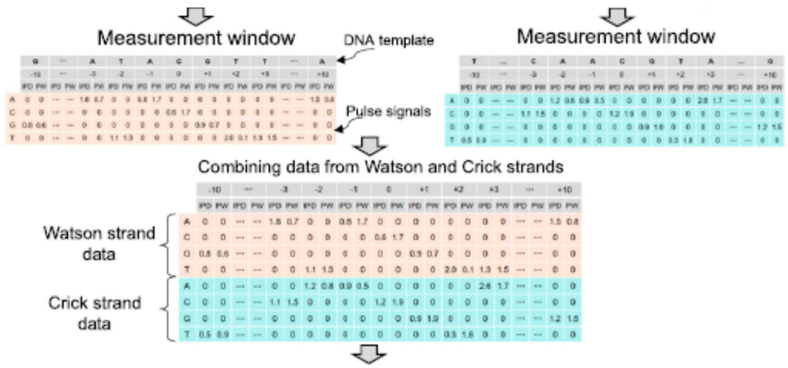
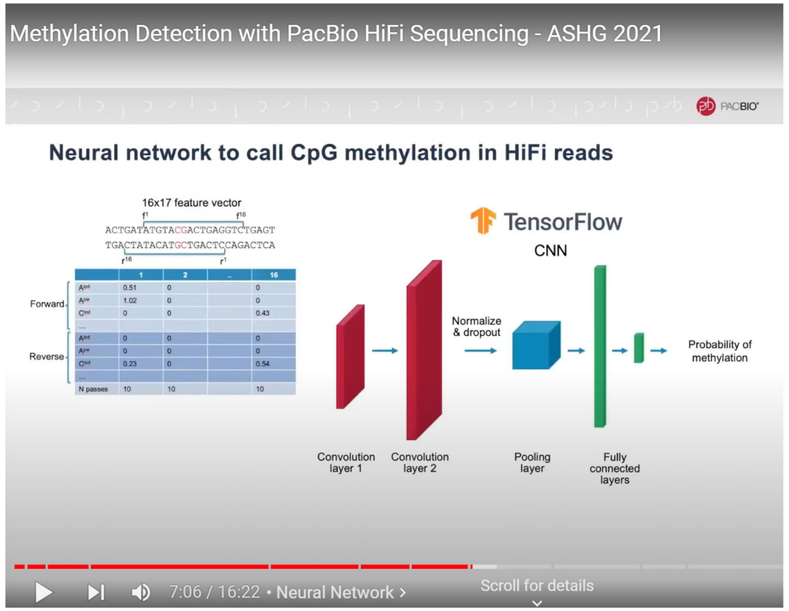
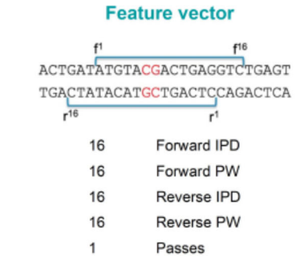
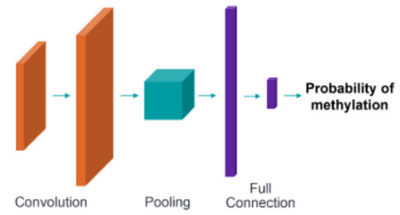
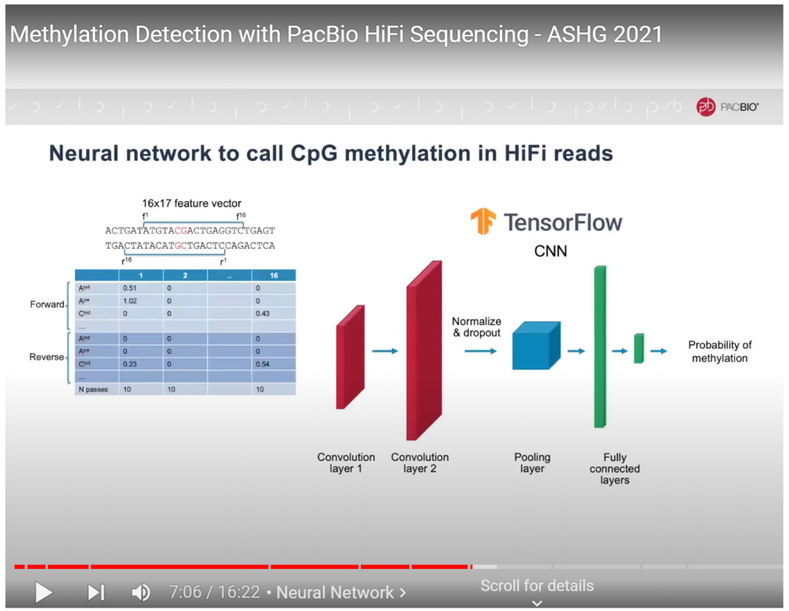
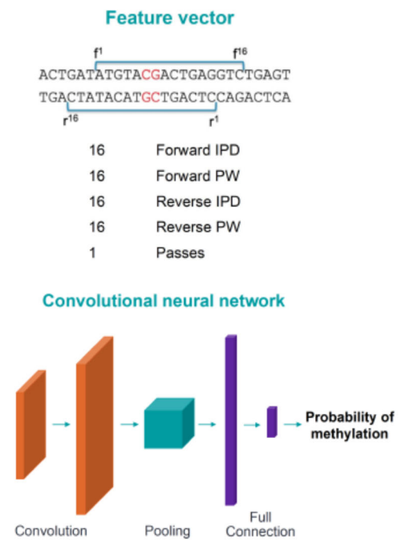
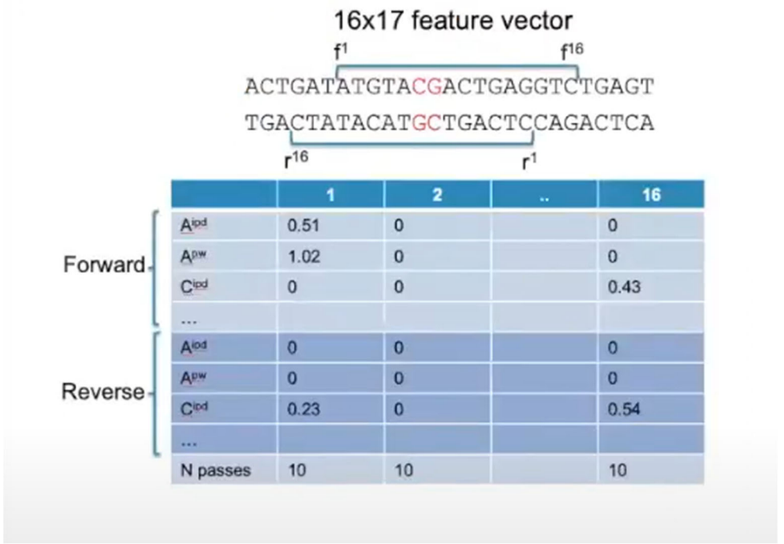


Figure 1. Kinetic signatures. Example trace showing pulse width (time of incorporation) and interpulse duration (time between adjacent incorporations). Image modified from Flusberg et al. (2010)[1].

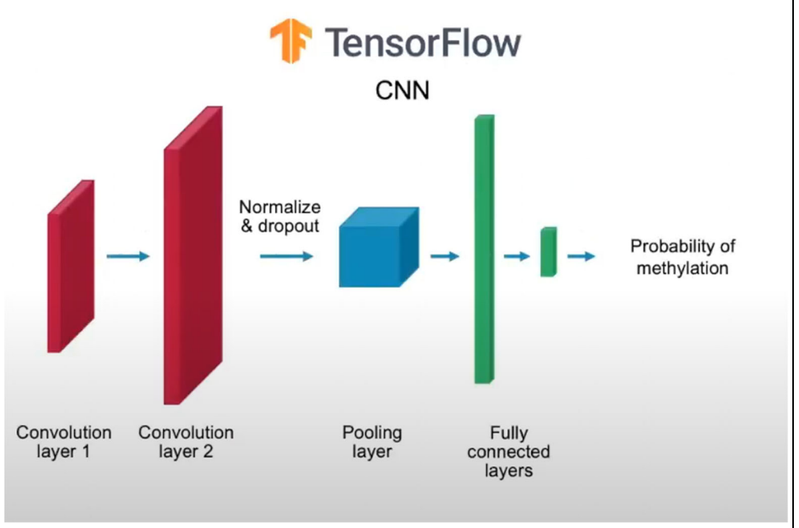
EX2047, Left Column.

(b) creating an input data structure, the input data structure comprising a window of the nucleotides sequenced in the sample nucleic acid molecule, wherein the input data structure includes, for each nucleotide within the window, the properties: the identity of the nucleotide, a position of the nucleotide with respect to a target position within the window, the width of the pulse corresponding to the nucleotide, and the interpulse duration;	 EX1030 at 2, Fig. 1 (partial) "For a cytosine subjected to methylation analysis, IPDs, PWs, and sequence context surrounding that cytosine were organized into a data matrix, referred to as a measurement window. For illustration purposes, the 10 nt upstream and downstream of the cytosine within a CpG site in question were presented as 5′-G[CCATGC]ATA***C***GTT[GATGCA]A-3′ for the Watson strand. The bases in the brackets were left out (denoted by "...") for the sake of simplicity. In this case, the measurement window size, including the interrogated cytosine in the middle, was 21 nt. For a position of -3 corresponding to the base of adenine ("A"), the IPD (1.8) and PW (0.7) associated with "A" were filled in the corresponding cells between a column of "-3" and a row of "A." The other cells in the same columns were filled by "0." The remaining IPDs and PWs related to the 21-nt sequence context were filled in that measurement window based on the same rule. The kinetic signals and sequence context originating from the Crick strand (′5-T[TTGCAT]CAA***C***GTA[TGCATG]G-3′) were also processed similarly. The measurement windows for two CpG sites complementary to each other (i.e., the Watson strand and the Crick strand) were combined for downstream analysis." EX1030 at 2, Fig. 1 legend, emphasis in the original. "We would hereby use the data processing of kinetic signals and sequence context from the Watson strand as an example. The position of an interrogated cytosine within a CpG site in a template DNA was denoted as position 0. For illustration purposes, the Watson and Crick strand templates comprising 10 nucleotides (nt) upstream and downstream of the cytosine in question were presented as 5′-G[CCATGC]ATA***C***GTT[GATGCA] A-3′ and 5′-T[TTGCAT]CAA***C***GTA[TGCATG]G-3′, respectively. The bases in the brackets were left out in Fig. 1 for the sake of simplicity. In this case, the measurement window size including the interrogated cytosine itself (in the center) was 21 nt. For the position of -3 corresponding to the base of adenine ("A"), the IPD (1.8) and PW (0.7) associated with "A" were filled in the intersection places (called cells) between a column of "-3" and a row of "A." The other cells between a	Dr. Wenger showed a 16x17 feature vector for the new tool under development in PacBio as an input data structure for "[n]eural network to call CpG methylation in HiFi reads." EX2024 screenshot at 7:06.  EX2024 screenshot at 7:06. "The basic idea is for every CpG site in a read, . . . to produce a feature vector that consists of the kinetics, the pulse widths and the interpulse durations in a 16 base window around that CpG site on both strands of the sequence. That's encoded in a vector and then fed into a convolutional neural network model implemented in TensorFlow and that outputs a probability of methylation for that CpG site under that assumption that the methylation is symmetric on the two strands of DNA." EX2024 at 7:22-8:02.	**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**  **Convolutional neural network**  **Figure 3. Primrose overview.** Visualization of the feature vector and neural network implemented in Primrose. EX2047, Middle Column.

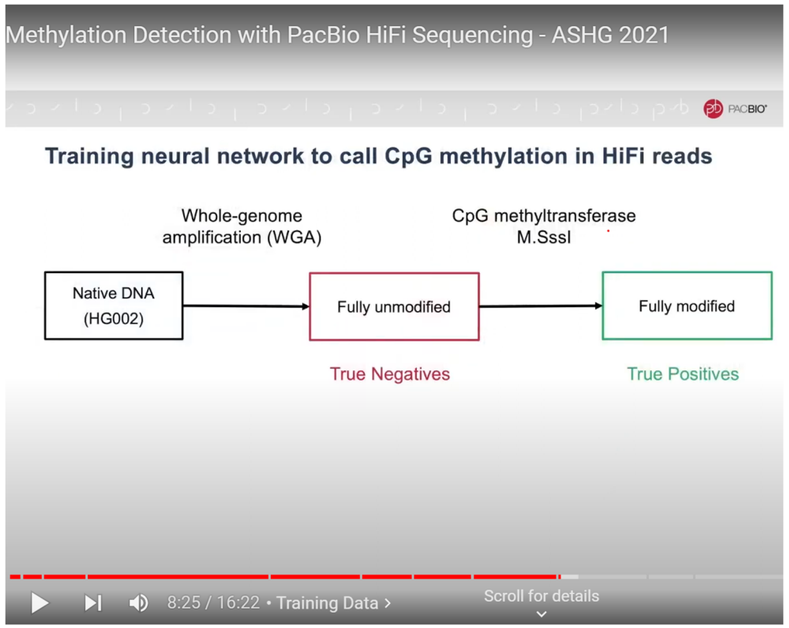
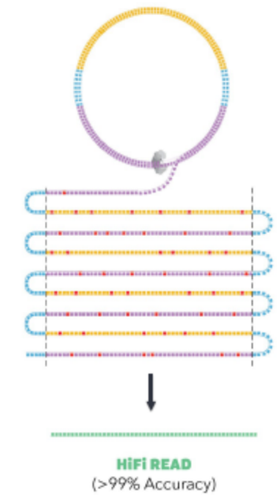
	column of -3 and rows of cytosine (“C”), guanine (“G”), and thymine (“T”) were filled by “0.” Other IPDs and PWs related to the 21-nt sequence context were filled in corresponding cells in that measurement window. The kinetic signals and sequence context originating from the Crick strand were similarly processed.” EX1030 at 3, emphasis in the original.		
(c) inputting the input data structure into a model,	“The measurement windows for two CpG sites complementary to each other (i.e., the Watson strand and the Crick strand) were combined for downstream analysis.” EX1030 at 2, Fig. 1 legend. “As nearly all methylated CpG sites would occur on both strands symmetrically (12), we combined the measurement window flanking a CpG site from the Watson strand with that flanking the paired CpG site from the Crick strand, forming a combined measurement window for downstream analysis.” EX1030 at 3.	Dr. Wenger Dr. Wenger presented feeding the feature vector as an input data structure to the TensorFlor CNN model to call CpG methylation. EX2024 screenshot at 7:06.  EX2024 screenshot at 7:06 (annotated). “The basic idea is for every CpG site in a read, . . . to produce a feature vector that consists of the kinetics, the pulse widths and the interpulse durations in a 16 base window around that CpG site on both strands of the sequence. That’s encoded in a vector and then fed into a convolutional neural network model implemented in TensorFlow and that outputs a probability of methylation for that CpG site under that assumption that the methylation is symmetric on the two strands of DNA.” EX2024 at 7:22-8:02.	**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.  **Figure 3. Primrose overview.** Visualization of the feature vector and neural network implemented in Primrose. EX2047, Middle Column.

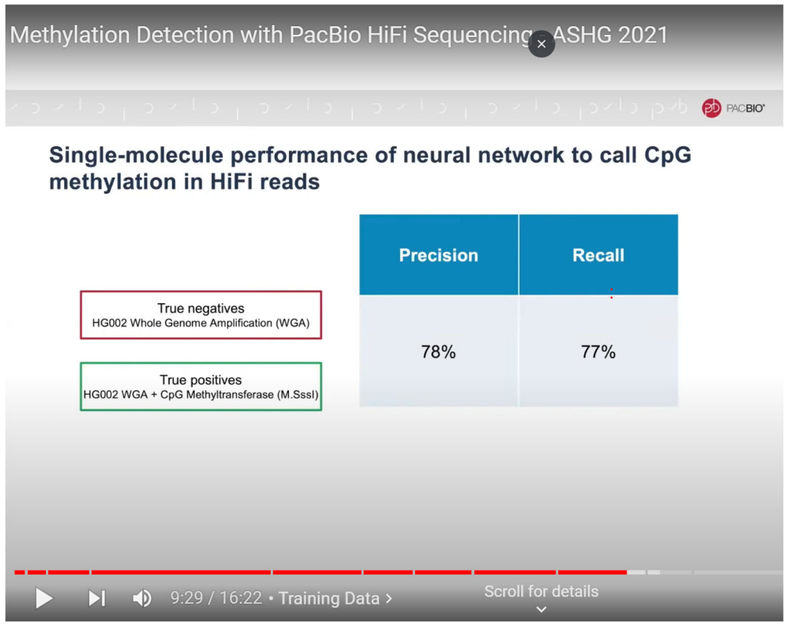


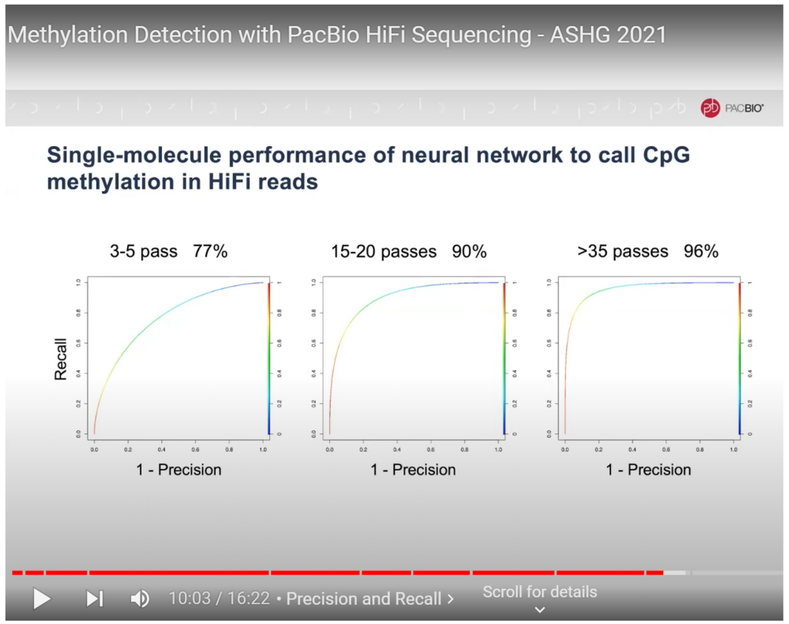
EX2024 screenshot at 7:06, 16x17 feature vector.



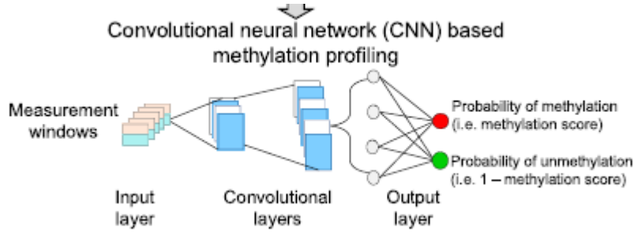
EX2024 screenshot at 7:06, TensorFlow CNN.

the model trained by: receiving a first plurality of first data structures, each first data structure of the first plurality of first data structures corresponding to a respective window of nucleotides sequenced in a respective nucleic acid molecule of a plurality of first nucleic acid molecules, wherein each of the first nucleic acid molecules is sequenced by measuring pulses in the optical signal corresponding to the nucleotides, wherein the modification has a known first state in a nucleotide at a target position in each window of each first nucleic acid molecule, each first data	"A number of combined measurement windows originating from methylated and unmethylated cytosines were used for training a CNN, so as to differentiate methylated and unmethylated cytosines in test samples. CNN involved input layer, convolutional layers, and output layer. The measurement windows were fed into the input layer, followed by the process of convolutional layers; then, the probability of methylation (range: 0 to 1) for a CpG was generated through the output layer based on a sigmoid function. This approach was referred to as the 'holistic kinetic (HK) model' (HK model)." EX1030 at 2, Fig. 1 legend. "We utilized a number of combined measurement windows originating from methylated and unmethylated cytosines, to train a CNN." EX1030 at 3. "The unmethylated dataset contained the sequencing results from amplified DNA that was prepared via whole genome amplification (WGA);" and "the methylated dataset contained the sequencing results from DNA treated by the M.SssI" that "rendered CpG sites methylated." EX1030 at 3-4. "[T]he model was trained and validated using SMRT sequencing datasets, including an unmethylated dataset (i.e., the negative dataset) and a methylated dataset (i.e., the positive dataset). The unmethylated dataset contained the sequencing results from amplified DNA that was prepared via whole genome amplification (WGA) (denoted as the WGA dataset). The use of unmodified nucleotides in the WGA resulted in the amplified DNA containing nearly no base modifications (with the exception of the small amount of input genomic DNA). The methylated dataset contained the sequencing results from DNA treated by the M.SssI (a CpG methyltransferase, isolated from a strain of *Escherichia coli* which contains the methyltransferase gene from *Sprioplasma* sp. strain MQ1, would methylate all CpG sites in a double-stranded DNA) prior to sequencing (denoted as the M.SssI-treated dataset). M.SssI methyltransferase rendered CpG sites methylated (13). Among the sequenced CpG sites within the dataset of the M.SssI-treated sample, half was used for training the HK model. Within the WGA dataset, an equal number of CpG sites were randomly sampled for training the HK model. The remaining half of the CpG sites within the dataset of the M.SssI-treated sample and the same number from the WGA dataset were used for validation of the model." EX1030 at 3-4. "For the Sequel I sequencing kit 3.0, we used 328,233 CpG sites from an M.SssI-treated DNA sample (fully methylated) and 328,233 CpG sites from a WGA sample (fully unmethylated) to train the HK model." EX1030 at 4.	 EX2024 screenshot at 8:25. "In addition to developing the feature for this model and the model structure, the other key attribute to producing a neural network model is training data. And for that you want examples of methylation where you have sequences that are either fully methylated or fully unmethylated. And in order to train this model we produced data in a manner where we start with native human DNA for the sample HG002 and in this native human DNA of course some of the sites are methylated and some are not. We then performed whole genome amplification using a PCR-based approach which produces then fully unmodified DNA. This library or a library was prepared from that DNA and sequenced and that serves as a large set genome-wide true negatives of DNA strands that are not modified. Then to produce fully modified sites we treated this whole genome amplified HG002 DNA with a CpG methyltransferase enzyme which efficiently adds methylation to any CpG site in the DNA and this then produces fully modified sequences which again were sequenced to produce a large set of true positives genome wide. These two sets were passed into the neural network to train the model. Some regions of the genome were held out to allow evaluation of precision and recall." EX2024 at 8:07-9:28.	**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 tc 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 tc 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. EX2047, Left Column.

structure comprising values for the same properties as the input data structure,			
storing a plurality of first training samples, each including one of the first plurality of first data structures and a first label indicating the first state of the nucleotide at the target position, and optimizing, using the plurality of first training samples, parameters of the model based on outputs of the model matching or not matching corresponding labels of the first labels when the first plurality of first data structures is input to the model, wherein an output of the	"Among the sequenced CpG sites within the dataset of the M.SssI-treated sample, half was used for training the HK model. Within the WGA dataset, an equal number of CpG sites were randomly sampled for training the HK model. The remaining half of the CpG sites within the dataset of the M.SssI-treated sample and the same number from the WGA dataset were used for validation of the model." EX1030 at 4. "To study how the window size of the measurement window and subread depth affected the performance of the HK model, we varied the measurement window sizes, covering 1, 3, 5, 7, 9, 11, 21, 31, 41, 51, and 61 nt. For a particular measurement window size, we further varied the subread depths, covering 1, 2, 3, 4, 5, 10, 15, 20, 25, and 30×. The HK model was first trained using a training dataset comparing 100,000 measurement windows each from the WGA and M.SssI-treated datasets. For each combination of window size and subread depth, we randomly sampled 2,000 CpG sites from a full dataset that did not overlap with the training dataset, thus forming a testing dataset." EX1030 at 4. "The measurement window size of 21 nt was a robust parameter for methylation analysis as such a window size appeared to reach a plateau value at a subread depth of 30× (*SI Appendix*, Fig. S7 *A* and *B*)." EX1030 at 5.	PacBio's new tool under development to call 5mC used some regions of the fully unmethylated sequences and fully methylated sequences for validation and optimization of the model by "evaluation of precision and recall." EX2024 screenshot at 9:24-9:27 ("some regions of the genome were held out to allow evaluation of precision and recall"); Koo declaration, EX2041.  EX2024 screenshot at 9:29. "Some regions of the genome were held out to allow evaluation of precision and recall." EX2024 9:24-9:28.	

model specifies whether the nucleotide at the target position in the respective window has the modification,		 EX2024 screenshot at 10:03. “Some regions of the genome were held out to allow evaluation of precision and recall and with the typical HiFi read lengths used in this study which were around 10 to 15,000 base pair reads, the precision and recall of CpG methylation at individual sites in individual reads was around 78 percent. The accuracy of calling methylation does depend on the number of passes of a of a HiFi read so how many serial observations of the molecule were made. With 10 to 15,000 base pair reads the typical number is somewhere around six to ten observations and you can see here as you as you start on the left showing the precision and recall trade-off with three to five passes of a molecule where the optimal balance is around 77 percent accuracy. As you move to the higher number of passes 15 to 20 passes the accuracy improves and at extremely high number of passes you get a nearly perfect calling.” EX2024 9:24-10:29.	

(d) determining, using the model, whether the modification is present in a nucleotide at the target position within the window in the input data structure.

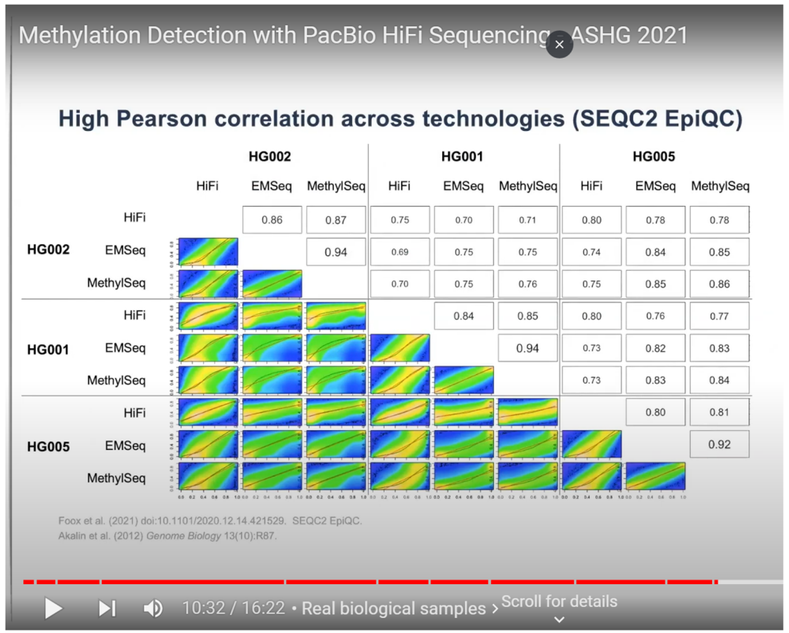


EX1030 at 2, Fig. 1 (partial), annotated.

"CNN involved input layer, convolutional layers, and output layer. The measurement windows were fed into the input layer, followed by the process of convolutional layers; then, the probability of methylation (range: 0 to 1) for a CpG was generated through the output layer based on a sigmoid function. This approach was referred to as the 'holistic kinetic (HK) model' (HK model)." EX1030 at 2, Fig. 1 legend.

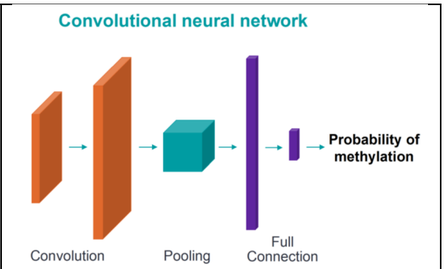
"**Methylation Determination Using the HK Model for Biological Samples.** To further validate whether the trained HK model could be used for analyzing real biological samples, we sequenced 11 tissue DNA samples using the Sequel II sequencer together with the Sequel II sequencing kit 1.0 (PacBio) (*SI Appendix*, Table S2). We obtained a median of 6 million sequenced molecules, with a median of 5.9 kilobases (kb) in size. The median subread depth was 4.3× (IQR: 3.6 to 6.7×). Each sample was also sequenced by BS-seq to a median of 50 million paired reads. The methylation states across CpG sites were determined by the Methy–Pipe software (14).

We compared the overall methylation levels between two measurements by the HK model and BS-seq. The overall methylation levels were defined as the percentage of CpG sites determined to be methylated among all sequenced CpG sites. Fig. 4 shows that the overall methylation levels across samples analyzed by the HK model correlated well with those quantified by BS-seq ($r = 0.99$; P value < 0.0001)." EX1030 at 6.



EX2024 screenshot at 10:32

"Moving beyond this synthetically constructed sequence in which or synthetically constructed set with either purely modified or purely unmodified sets we can also apply this model trained in that context to real biological samples and this has been done in work by the SEQC2 consortium EpiQC working group in which they applied PacBio HiFi sequencing and other technologies for detecting methylation to three different reference human materials from genome in a bottle so this is the HG002 sample and also HG001 and HG005. These are three different all three different individuals of different backgrounds and for all these individuals there's a high coverage HiFi sequencing in addition to EMSeq and MethylSeq which are short read sequencing-based approaches to detecting methylation with MethylSeq being bisulfite conversion. What we see is that in these real biological contexts you see a high correlation between the methylation calls of HiFi and the other technologies, importantly seeing that the correlation of HiFi to, for example MethylSeq, is higher within an individual than is HiFi-HiFi across individuals." EX2024 10:30-11:58.



EX2047, Middle Column.

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).

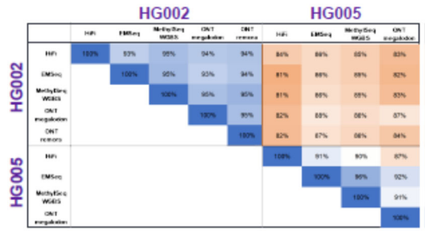


Figure 4. Technology comparison. Pearson correlation by position, compared across technologies and GIAB samples. HiFi datasets were ~30x depth of coverage per sample.

EX2047, Middle Column.

Conclusions

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

EX2047, Right Column.