

Petitioner's Demonstratives

Pacific Biosciences of California, Inc. v. Chinese University of Hong Kong

IPR2024-00028

U.S. Patent Nos. 11,091,794

Instituted Grounds

Claims	35 U.S.C. §	Prior Art
1-2, 7, 10-12, 14-15, 17-19	103	Flusberg, Flusberg-338, PacBio-BAM
3-6, 9	103	Flusberg, Flusberg-338, PacBio-BAM, Clement
8, 15-16	103	Flusberg, Flusberg-338, PacBio-BAM, PacBio Template Guide
13, 15-16	103	Flusberg, Flusberg-338, PacBio-BAM, Tsai

The '794 Patent

(54) DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

(57) ABSTRACT

Systems and methods for using **determination of base modification** in analyzing nucleic acid molecules and acquiring data for analysis of nucleic acid molecules are described herein. Base modifications may include methylations. Methods to determine base modifications may include using features derived from sequencing. These features may include the pulse width of an optical signal from sequencing bases, the interpulse duration of bases, and the identity of the bases. **Machine learning models** can be trained to detect the base modifications using these features. The relative modification or methylation levels between haplotypes may indicate a disorder. Modification or methylation statuses may also be used to detect chimeric molecules.

pre-treatment (e.g. digestion with restriction enzymes) may serve to enhance aspects of the invention (e.g. allowing the enrichment of CpG sites for analysis). The embodiments present in this disclosure could be used for detecting different types of **base modification**, for example, including but not limited to 4mC, 5hmC, 5fC, and 5caC, 1 mA, 3 mA, 7 mA, 3mC, 2mG, 6mG, 7mG, 3mT, and 4mT, etc. Such

site and "M" indicates a modified site. One example of base modification that is not based on methylation is oxidation changes, such as in 8-oxo-guanine.

EX1001, [54], [57], 2:44-48, 27:39-51

The '794 Patent

training. For simplicity, FIG. 9 shows only a statistical model developed by training, but any model or data analysis framework can be developed. Example data analysis frameworks include machine learning models, statistical models, and mathematical models. The statistical models could include, but was not limited to, linear regression, logistic regression, deep recurrent neural network (e.g., long short-term memory, LSTM), Bayes classifier, hidden Markov model (HMM), linear discriminant analysis (LDA), k-means clustering, density-based spatial clustering of applications with noise (DBSCAN), random forest algorithm, and support vector machine (SVM). A DNA stretch surrounding a base that was subjected to base modification analysis could be X-nt upstream and Y-nt downstream of that base, namely "measurement window."

Embodiments of the present invention can be used for, but is not limited to, single molecule sequencing. One type of single molecule sequencing is single molecule, real-time sequencing in which the progress of the sequencing of a single DNA molecule is monitored in real-time. One type of single molecule, real-time sequencing is that commercialized by Pacific Biosciences using their Single Molecule, Real-Time (SMRT) system. Methods may use the pulse width of a signal from sequencing bases, the interpulse duration (IPD) of bases, and the identity of the bases in order to detect a modification in a base or in a neighboring base. Another single molecule system is that based on nanopore sequencing. One example of a nanopore sequencing system is that commercialized by Oxford Nanopore Technologies.

The term "machine learning models" may include models based on using sample data (e.g., training data) to make predictions on test data, and thus may include supervised learning. Machine learning models often are developed using a computer or a processor. Machine learning models may include statistical models.

The model may include a supervised learning model. Supervised learning models may include different approaches and algorithms including analytical learning, artificial neural network, backpropagation, boosting (meta-algorithm), Bayesian statistics, case-based reasoning, decision tree learning, inductive logic programming, Gaussian process regression, genetic programming, group method of data handling, kernel estimators, learning automata, learning classifier systems, minimum message length (decision trees, decision graphs, etc.), multilinear subspace learning, naive Bayes classifier, maximum entropy classifier, conditional random field, Nearest Neighbor Algorithm, probably approximately correct learning (PAC) learning, ripple down rules, a knowledge acquisition methodology, symbolic machine learning algorithms, subsymbolic machine learning algorithms, support vector machines, Minimum Complexity Machines (MCM), random forests, ensembles of classifiers, ordinal classification, data pre-processing, handling imbalanced datasets, statistical relational learning, or Proaftn, a multicriteria classification algorithm. The model may linear regression, logistic regression, deep recurrent neural network (e.g., long short term memory, LSTM), Bayes classifier, hidden Markov model (HMM), linear discriminant analysis (LDA), k-means clustering, density-based spatial clustering of applications with noise (DBSCAN), random forest algorithm, support vector machine (SVM), or any model described herein.

EX1001, 2:54-67, 16:35-40, 26:34-48, 89:25-51

'794 Patent Claim 1

1. A method for detecting a modification of a nucleotide in a nucleic acid molecule, the method comprising:

(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;

(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;

(c) inputting the input data structure into a model, the model trained by:

- receiving a first plurality of first data structures, each first data structure of the first plurality of 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 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 model specifies whether the nucleotide at the target position in the respective window has the modification,

(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.

EX1001, claim 1

'794 Patent Claim 11

11. The method of claim 1, wherein the model includes a machine learning model, a principal component analysis, a convolutional neural network, or a logistic regression.

Dr. Koo:

Q. Okay, but sticking with Claim 1, Claim 1 recites the term "model," correct?

A. Claim 1, yes.

Q. Okay, but it doesn't recite "machine learning model"?

A. No. That comes in at 11.

Q. Must the model of Claim 1 be a convolutional neural network?

A. From my understanding, it does not have to be limited to a convolutional neural network model.

EX1001, claim 11; EX1049, 41:19-21, 78:9-12

Petition: Proposed Combination

Flusberg Publications disclose that nucleic acid molecules are sequenced using PacBio SMRT sequencing, generating pulse traces in which each nucleotide produces a pulse from which characteristic features may be derived. EX1006, 462-63, Fig. 1; EX1007, 23:14-22, 64:29-43, 66:46-67:4, Figs. 1 & 2. These features include the identities of each nucleotide in the sequence, as well as parameters including PW and IPD for each nucleotide. *Id.* Indeed, the Office recently

The Flusberg Publications teach that pulse features (e.g., PW and IPD) are tabulated for a window of nucleotides and used in a previously trained model to determine whether a nucleotide is methylated. *E.g.*, EX1006, 463, 465-466, Fig. 3;

To the extent the Flusberg Publications are not seen to expressly disclose each of nucleotide identity, position, PW, and IPD in an input data structure, this is taught by PacBio-BAM, which discloses the standard data structure in which PacBio SMRT sequencers store data. EX1011, 1-2. These data structures contain, *inter alia*, identity, position, PW, and IPD values for each nucleotide in a sequence. EX1011, 1-2, 4-5, 8; EX1012, 15; EX1002, ¶¶78-81.

Pet., 27-29; *see also id.*, 33-51

Petition: Rationale to Combine

A POSA would have had good reason to combine the Flusberg Publications with PacBio-BAM to achieve what is claimed with a reasonable expectation of success. The Flusberg Publications disclose detecting modifications using PacBio SMRT sequencing. PacBio-BAM teaches a BAM file containing values for all claimed properties that was generated in the course of SMRT sequencing. As established herein, BAM files generated by the SMRT sequencing platform contain data for the identity and position of bases along with PW and IPD. EX1011, 4-5, 8. PacBio-BAM further describes the file as compatible with “base modification detection.” EX1011, 8; *see* EX1021, 1-2 (using PacBio’s BAM file as an input data

Additionally, it was well-known and documented in the prior art that values for the claimed SMRT sequencing properties were routinely generated and known as useful in base modification detection. Specifically, POSAs appreciated that the sequence context (e.g., identity and positions of the bases in the vicinity of a modified base) strongly affects polymerase kinetics, and as a result, PW and IPD values. EX1006, 462-463; EX1007, 62:29-31; EX1008, 1-2; EX1009, 124;

43, 79:2-5, Figs. 18-19; EX1008, 2; EX1010, 130; EX1016, 3. Because of the documented interdependence of local sequence context and base modifications on PW and IPD, POSAs appreciated that local sequence effects should be considered when predicting base modifications. *E.g.*, EX1006, 462-63 (“[K]inetic signatures

Pet., 29-33; *see also id.*, 34-51

Flusberg

Fluorescence pulses in SMRT sequencing are characterized not only by their emission spectra but also by their duration and by the interval between successive pulses²⁵. These metrics, defined here as pulse width and interpulse duration (IPD), add valuable information about DNA polymerase kinetics. Pulse width is a function of all kinetic steps after nucleotide binding and up to fluorophore release, and IPD is determined by the kinetics of nucleotide binding and polymerase translocation²⁷. We have previously demonstrated that SMRT sequencing polymerase synthesis rates are sensitive to DNA primary and secondary structure²⁵. Therefore, we hypothesized that methylated bases

We describe the direct detection of DNA methylation, without bisulfite conversion, through single-molecule, real-time (SMRT) sequencing. In SMRT sequencing, DNA polymerases catalyze the incorporation of fluorescently labeled nucleotides into complementary nucleic acid strands. The arrival times and durations of the resulting fluorescence pulses yield information about polymerase kinetics and allow direct detection of modified nucleotides in the DNA template, including N6-methyladenine, 5-methylcytosine and 5-hydroxymethylcytosine. Measurement

polymerase active site²⁹. Consistent with this model, the kinetic impact of methylation is not restricted to the nucleotide incorporation opposite the modified base. Additionally, the IPD ratio patterns differed for the two methylated positions. As the only source of difference between these two loci in all three template types was the local sequence context (template sequences are listed in Supplementary Note 1), we conclude that, in general, the kinetic signatures of methylation will be sequence context-dependent.

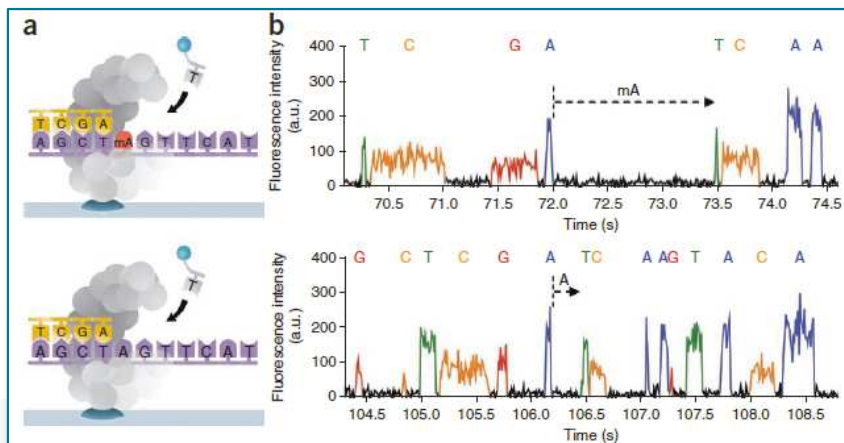


Figure 1 | Principle and corresponding example of detecting DNA methylation during SMRT sequencing. (a) Schematics of polymerase synthesis of DNA strands containing a methylated (top) or unmethylated (bottom) adenine. (b) Typical SMRT sequencing fluorescence traces for samples in a. Letters above the fluorescence trace pulses indicate the identity of the nucleotide incorporated into the growing complementary strand. Dashed arrows indicate the IPD before incorporation of the cognate thymine. For this typical example, the IPD is about five times larger for mA in the template compared to adenine.

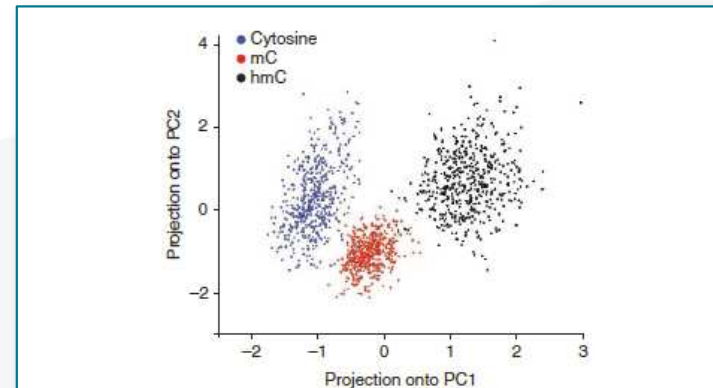


Figure 3 | Principal component analysis of cytosine, mC and hmC IPD and pulse width signatures. Each principal component is a linear combination of the mean IPD and pulse width at positions 71–79, which surround the variably modified template position 73. The weightings of IPD and pulse width at each position are shown in Supplementary Table 1. Data points on the plot were computed by projecting a random 20% subsample of the IPD and pulse width values onto the first two principal components (PC1 and PC2) and then converting to a z score.

EX1006, Abstract, 462-63, 465, Figs. 1, 3

Flusberg-338

(54) METHODS FOR IDENTIFYING NUCLEIC ACID MODIFICATIONS

Generally speaking, the methods of the invention involve monitoring of an analytical reaction to collect “reaction data,” wherein the reaction data is indicative of the progress of the reaction. Reaction data includes data collected directly from the reaction, as well as the results of various manipulations of that directly collected data, any or a combination of which can serve as a signal for the presence of a modification in the template nucleic acid. For example, certain types of reaction data are collected in real time during the course of the reaction, such as metrics related to reaction kinetics, affinity, rate, processivity, signal characteristics, and the like. As used herein, “kinetics,” “kinetic signature,” “kinetic response,” “activity,” and “behavior” of an enzyme (or other reaction component, or the reaction as a whole) generally refer to reaction data related to the function/progress of the enzyme (or component or reaction) under investigation and are often used interchangeably herein. Signal characteristics vary depending on the type of analytical reaction being monitored. For example, some reactions use detectable labels to tag one or more reaction components, and signal characteristics for a detectable label include, but are not limited to, the type of signal (e.g., wavelength, charge, etc.) and the shape of the signal (e.g., height, width, curve, etc.). Further, signal characteristics for multiple signals (e.g., temporally adjacent signals) can also be used, including, e.g., the distance between signals during a reaction, the number of extra signals (e.g., that do not correspond to the progress of the reaction), internal complementarity, and the local signal context (i.e., one or more signal that precede and/or follow a given signal). For

(57)

ABSTRACT

Methods, compositions, and systems are provided for characterization of modified nucleic acids. In certain preferred embodiments, single molecule sequencing methods are provided for identification of modified nucleotides within nucleic acid sequences. Modifications detectable by the methods provided herein include chemically modified bases, enzymatically modified bases, abasic sites, non-natural bases, secondary structures, and agents bound to a template nucleic acid.

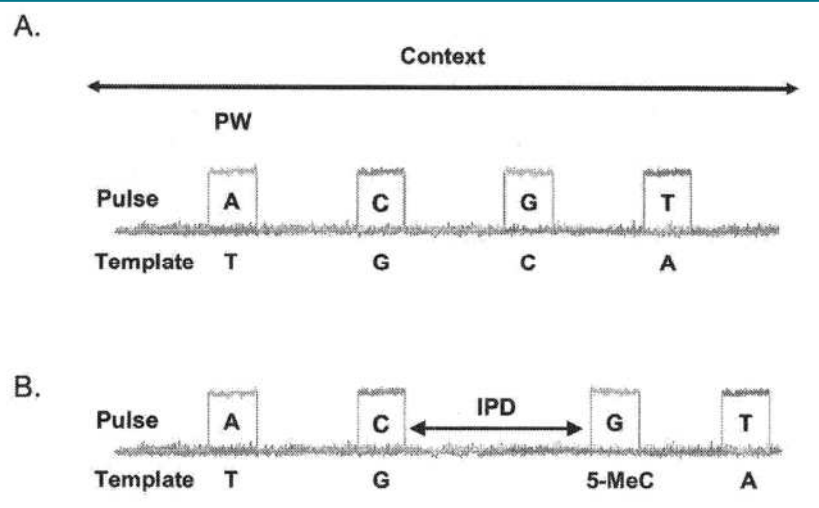


FIG. 2A and FIG. 2B provide illustrative examples of various types of reaction data in the context of a pulse trace.

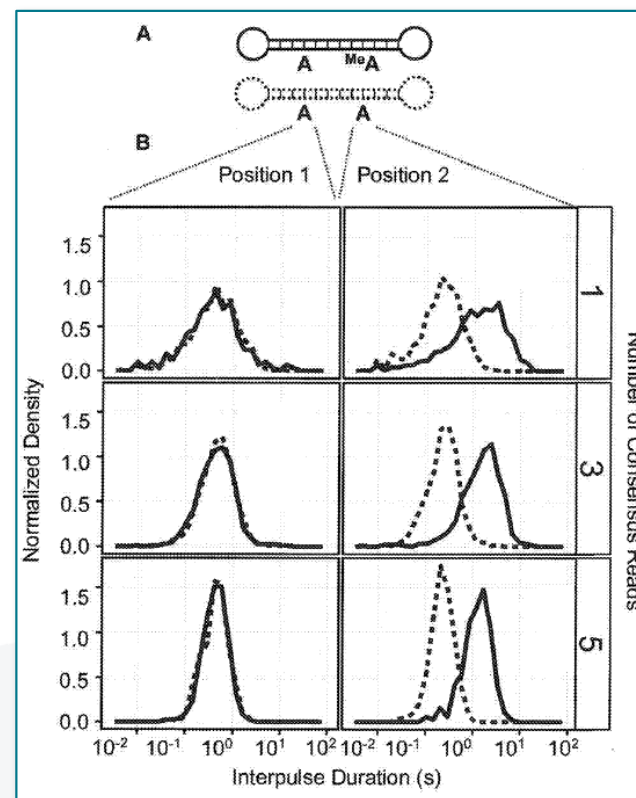
Flusberg-338

entirety for all purposes. A first stage of this approach includes the development of a classifier for distinguishing methylated from unmethylated cytosines in a nucleic acid template. A set of features is measured for every pulse (discrete event or signal) in the trace. For example, a set of measurable features might be {pulse width, pulse duration, pulse height, pulse amplitude variability}. Call the values of these features $f_1^i, f_2^i, \dots, f_k^i$, for pulse i . For each cytosine pulse in the methylated and unmethylated template data sets (which may or may not be restricted to CpG), tabulation is performed for the local pulse features

$$\vec{f} = \{f_1^{i-3}, f_2^{i-3}, \dots, f_k^{i-3}, f_1^i, f_2^i, \dots, f_k^i, \dots, f_1^{i+3}, f_2^{i+3}, \dots, f_k^{i+3}\}.$$

In this example, a local context extending 3 pulses to the left and right of the pulse of interest is assumed, but this context size is flexible and, in certain embodiments, can be application-specific.

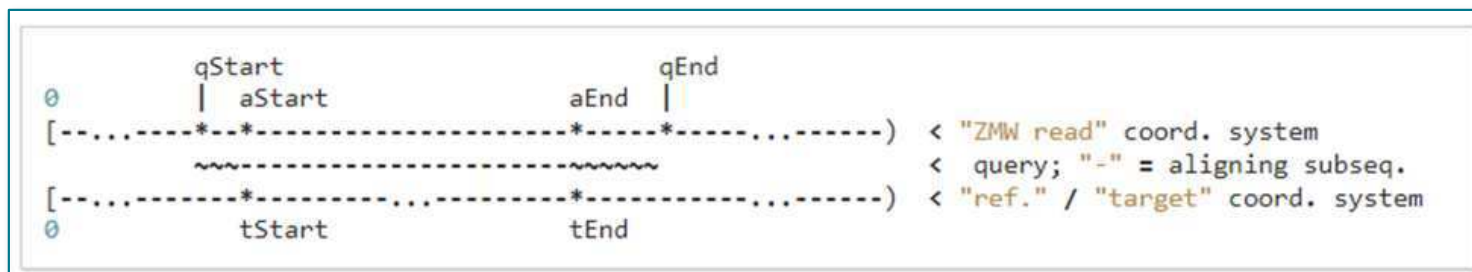
measured spectrum. Many different statistical classification algorithms known in the art can be used in this classifier. Certain preferred algorithms include classification-and-regression trees (CART), naïve Bayesian classifiers, kernel density methods, linear discriminant functions, and neural networks. Further, the pulse classifier does not need to be



deoxyribonucleotides.”). Even where presenting experimental examples using IPD, Flusberg-338 expressly states additional metrics are not excluded. EX1007, 67:64-67 (“it will be clear to one of ordinary skill that the approach can be extended to pulse metrics other than IPD, such as pulse width...etc.”). Nor does PO address

EX1007, 63:22-27, 66:54-67:4, Fig. 14; Reply, 14

PacBio-BAM



BAM format specification for PacBio

The BAM format is a binary, compressed, record-oriented container format for **raw or aligned sequence reads**. The associated SAM format is a text representation of the same data. The **specifications for BAM/SAM** are maintained by the SAM/BAM Format Specification Working Group.

PacBio-produced BAM files are fully **compatible with the BAM specification**. In this document we describe the way we make use of the extensibility mechanisms of the BAM specification to encode PacBio-specific information, as well as conventions we adhere to.

Encoding of kinetics pulse features

Interpulse duration (IPD) and pulsewidth are measured in frames; natively they are recorded as a `uint16` per pulse/base event. They may be **encoded in BAM read tags** in one of two fashions:

- losslessly as an array of `uint16`; necessary for PacBio-internal applications but entails greater disk space usage.
- lossy 8-bit compression stored as a `uint8` array, following the codec specified below ("codec V1"). Provides a substantial disk-space savings without affecting important production use cases (**base modification detection**).

Petition—Corroborating Prior Art

Additionally, it was well-known and documented in the prior art that values for the claimed SMRT sequencing properties were routinely generated and known as useful in base modification detection. Specifically, POSAs appreciated that the sequence context (e.g., identity and positions of the bases in the vicinity of a modified base) strongly affects polymerase kinetics, and as a result, PW and IPD values. EX1006, 462-463; EX1007, 62:29-31; EX1008, 1-2; EX1009, 124; EX1016, 2; EX1010, 130. POSAs.

Pet. 30.

Moreover, as Dr. Dessimoz explains, the scientific literature at the time further corroborates the obviousness of considering data for nucleotide identity, position, PW, and IPD together.

Pet. 31.

EX1021, 1-2 (using PacBio's BAM file as an input data structure for base modification detection)

Pet. 29.

Corroborating Prior Art: PacBio White Paper

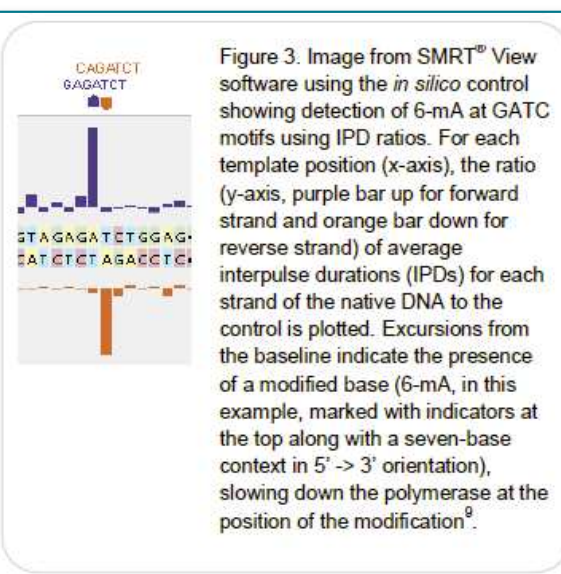
Detecting DNA Base Modifications Using Single Molecule, Real-Time Sequencing

- Teaches effect of local context over window on modification signature and describes typical experiment that takes kinetic properties and sequence context into account

using the *in silico* control. Future commercial products are expected to improve use of the sequence context to better identify and interpret modification signals.

Based on the measurement and analysis principle outlined above, a typical experiment for detecting DNA base modifications can be designed as follows:

- Obtain native DNA of interest and prepare a SMRTbell™ library for SMRT sequencing (see the Pacific Biosciences Sample Preparation and Sequencing Guide).
- If using an amplified control design, use a small aliquot of the original DNA sample and perform a whole-genome amplification (WGA) reaction to obtain the control DNA sample lacking any base modifications. Then prepare a SMRTbell library for this sample.
- Perform adequate SMRT sequencing (for both samples when using an amplified control) to obtain the necessary coverage needed for the modifications under study, and sufficient overall coverage to characterize the genome (or portion of the genome) of interest.
- Use the bioinformatics tools to perform kinetic analysis for base modification⁹.



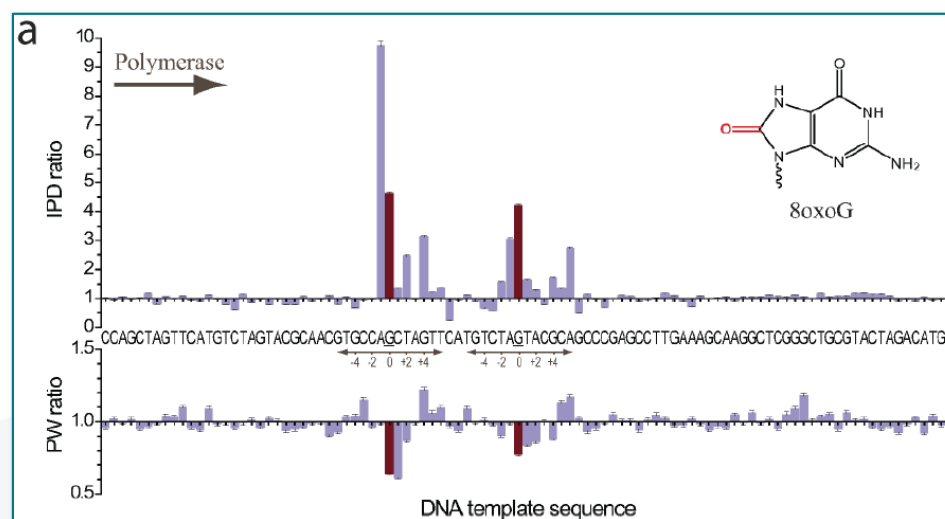
Kinetic effects are not necessarily limited to just the nucleotide incorporation opposite the modified base position in the DNA template. This is because the DNA polymerase is in intimate contact with the DNA over an extended region of approximately twelve bases, and the modified base can impart effects on the polymerase dynamics at several positions over this region. This results in kinetic "signatures" which can aid in identifying the type of base modification.

Corroborating Prior Art: Clark

Direct Detection and Sequencing of Damaged DNA Bases

- Teaches both PW and IPD are principal pulse metrics and that their signature is dependent on local sequence context, as visible in a window around the modification

Recently, single-molecule, real-time (SMRT) DNA sequencing has been described for the direct detection of methylated and hydroxymethylated DNA bases [5]. In SMRT sequencing, the progression of single molecules of DNA polymerase is monitored in real time during base incorporations using fluorescent phospholinked nucleotides [6,7]. The dynamics of DNA polymerization is thereby recorded in the form of a train of fluorescent pulses. The length of time that the polymerase retains a nucleotide bound in its active site (**pulse width, PW**), and the time interval between successive nucleotide-bound states (**interpulse duration, IPD**) are the **principal pulse metrics** used in the analysis to ascertain that the polymerase kinetics was altered in the modification-containing template when compared to an unmodified control template [5]. Because the DNA polymerase is in contact with the modified base over a region of ~11 bases [8], the kinetic effects are not necessarily restricted to the nucleotide incorporation opposite of the modified base. The magnitude and extent of the kinetic signature is dependent on the type of the base modification and the local sequence context. This results in distinct, more complex kinetic signatures that can be used to discriminate between different chemical base modifications [5].



EX1008, 1-2, Fig. 1; Pet., 7-8, 23-24, 30; Reply, 12-13, 15

Corroborating Prior Art: *kineticsTools*

- Detects methylation modifications using a trained model

The **modification detection tool** provides results in a variety of formats suitable for in-depth statistical analysis, quick reference, and consumption by visualization tools such as PacBio SMRTView. Results are generally indexed by reference position and reference strand. In all cases the strand value refers to the strand carrying the modification in DNA sample. Remember that the kinetic effect of the modification is observed in read sequences aligning to the opposite strand. So reads aligning to the positive strand carry information about modification on the negative strand and vice versa, but in this toolkit we always report the strand containing the putative modification.

The modifications.gff is compliant with the GFF Version 3 specification (<http://www.sequenceontology.org/gff3.shtml>). Each template position / strand pair whose p-value exceeds the pvalue threshold appears as a row. The template position is 1-based, per the GFF spec. **The strand column refers to the strand carrying the detected modification**, which is the opposite strand from those used to detect the modification. The GFF confidence column is a Phred-transformed pvalue of detection.

- Trained model uses BAM file (containing basecalls, IPD, and PW) as input data structure

Basic use with BAM input, GFF+HDF5 output:

```
ipdSummary aligned.bam --reference ref.fasta --identify m6A,m4C --gff basemods.gff --csv_h5 kinetics.h5
```

- Detects modifications in a window of nucleotides

Studies of the relationship between IPD and sequence context reveal that most of the variation in mean IPD across a genome can be predicted from a **12-base sequence context** surrounding the active site of the DNA polymerase. The **bounds of the relevant context window correspond to the window of DNA in contact with the polymerase**, as seen in DNA/polymerase crystal structures. To simplify the process of finding DNA modifications with PacBio data, the tool includes a **pre-trained lookup table mapping 12-mer DNA sequences** to mean IPDs observed in C2 chemistry.

EX1021, 1-3; Pet., 9, 22, 24, 29, 45; Reply, 4, 12

Corroborating Prior Art: Tahir

Specifically, the prior art at the time expressly embraced neural networks for detecting methylation status including the Keras framework used by PO. EX1017, 16 (cataloging deep learning resources, including Keras, for genomics applications); EX1035, 3, (“implemented by Keras framework”), 5 (describing “an intelligent and robust method” for methylation detection); EX1022, Fig. 1 (illustrating neural network for detecting methylation in ONT data); see EX1030, 10 (“The program for the CNN model was implemented on the basis of the Keras deep learning framework (<https://keras.io/>).”); EX1002, ¶¶87-90..

Pet. 33.

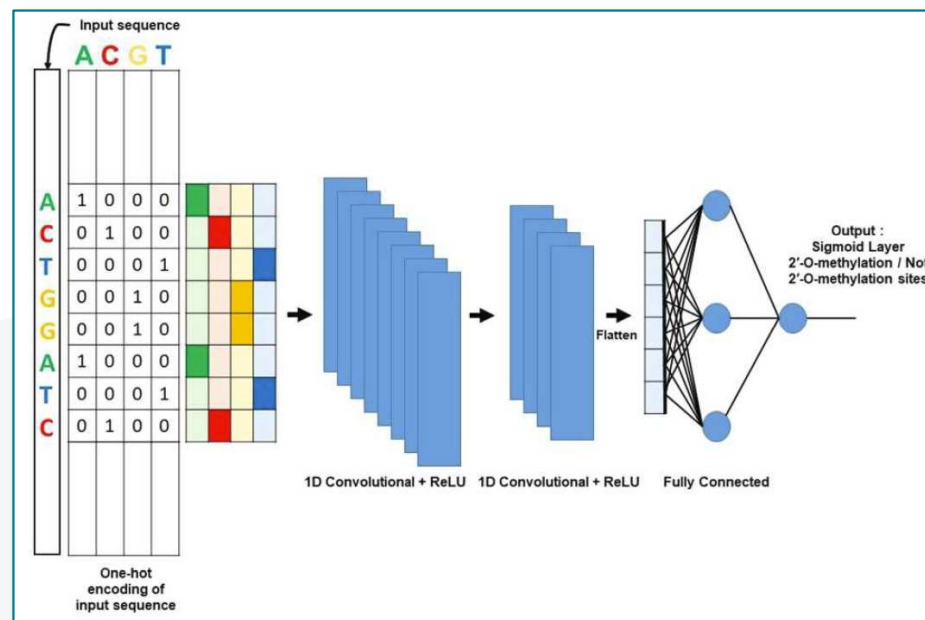
Tahir:

The sigmoid layer that makes predictions whether a given sequence is a 2'-O-methylation sites or not 2'-O-methylation sites.

$$\text{Sigmoid}(y) = \frac{1}{1 + e^{-y}} \quad (6)$$

The sigmoid () function outputs normalized class probabilities for a given input. The output of this layer is scaled to the [0, 1] by sigmoid function.

In this study, the proposed model is implemented by Keras framework (Chollet, 2015). Adam optimizer is used for optimization and the learning rate is set to 0.0001. Batch size is set to 10. Number of Epochs is set to 50.



EX1035, 3-4, Fig. 1; Pet., 33; Reply, 7 n.1; 22-23; EX1002, ¶¶26, 43-44, 88-89, 138, 155

PO Fails to Rebut Obviousness

- Prior art teaches claimed “input data structure”
- No teaching away from including nucleotide identity
- No teaching away from including pulse width
- No secondary indicia attributable to claimed invention

I.	INTRODUCTION.....	1
II.	THE '794 PATENT.....	4
A.	Claim 1	5
B.	Prosecution of Claim 1	7
C.	Claim Construction.....	11
D.	Person of Ordinary Skill in the Art	11
III.	CLAIM 1 IS NOT OBVIOUS OVER GROUND 1.....	11
A.	Flusberg, the '338 Patent and the PacBio-BAM do not Disclose or Render Obvious the Combination of Nucleotide Identity, Position, PW, IPD in an Input Data Structure of Claim 1.	12
1.	Identity of the nucleotide	14
a.	No reference includes “identity of nucleotide” with any other property recited in claim 1.....	17
b.	The Petition’s other listed reason do not support unpatentability under Ground 1.....	21
c.	“Sequence context” does not mean “identity of the nucleotide”	29
2.	A POSA would not have included pulse width with the other properties of the claimed input data structure.	31
IV.	OBJECTIVE INDICIA OF NONOBVIOUSNESS SHOW THAT THE INPUT DATA STRUCTURE OF CLAIM 1 WAS NOT OBVIOUS AT THE TIME OF THE INVENTION.	36
A.	The Method of Claim 1 Solved the Long-Felt Unsolved Need for Direct Detection of DNA Modifications by SMRT Sequencing Without Bisulfite Conversion.....	37

See generally POR; Reply

PO Fails to Distinguish BAM File from Claim 1(b)

PO's Arguments Not Commensurate with Scope of Claims

PO's argument:

PO's "input data structure" arguments also fail because they are not commensurate with the broad scope of the claims. For example, Dr. Koo testified that, in computer science, data structures are simply "objects that contain [] information." EX1049, 132:21-25. PO argues the BAM file cannot be directly read by a trainable model, but the claims recite no such requirement. PO's claims generically recite a "model," and the specification confirms the staggering breadth of the term "model." PO does not propose a limiting claim construction for "model." Nor could it. The patent places essentially no restriction on models that may be used within the scope of the invention and expressly permits inclusion of sub-modules.

Dr. Koo:

Q. Okay. The Element 1B doesn't recite the term "consisting of"; am I correct in that?

A. The input data structure -- I do not see the words "consisting of."

Q. Well, I'm just asking if the term "compressed" is there.

A. The term "compressed" is not here. It's not present in this claim (indicating).

Q. Okay, and the term "decompressed" is also not present in this claim, correct?

A. Both of those terms are not present in this claim -- in -- in 1B.

PO Fails to Distinguish BAM File from Claim 1(b)

PO Advances the same Argument Rejected at Institution

Reply:

PO argues (p. 20) that PacBio's BAM file cannot satisfy the input data structure of Element 1.2. According to PO and Dr. Koo, the BAM file is a compressed binary format and "it comprises text data, not an input data array of numerical values" and "includes extraneous information...that needs to be parsed out before the data could be used by a trainable model." EX2041, ¶¶129-130. This argument is not persuasive and fails to undermine any of the Board's sound reasons for rejecting this same argument at institution. DI, 21-22.

Institution decision:

Patent Owner does not explain why a person of ordinary skill in the art could not have done so using the substantively identical teaching in the asserted PacBio-BAM.⁷ Third, we are not persuaded by Patent Owner's implication that a skilled artisan would thoughtlessly incorporate "text data" that would sabotage the utility of an input data structure, as Patent Owner suggests would occur. Prelim. Resp. 31; see *KSR Int'l*, 550 U.S. at 421 ("A person of ordinary skill is also a person of ordinary creativity, not an automaton.").

Reply, 3; DI, 21-22

Dr. Dessimoz Confirms Inputting BAM Data to Model Trivial

Declaration testimony:

and position of bases along with IPD and PW. EX1011, 4-5, 8. PacBio-BAM further describes the file as compatible with base modification detection. EX1011, 8 (“Provides a substantial disk-space savings without affecting important production use cases (base modification detection).”); *see also* EX1021, 1-2 (using PacBio’s BAM file as an input data structure for base modification detection). PacBio-BAM would thus have represented a preexisting, off-the-shelf option for a data structure containing the different variables known to be most important in detecting base modifications, and a person of ordinary skill in the art would have

Cross examination confirms:

For instance, we do not need to go and retrieve additional information that would not be contained in the BAM file. It's just a straightforward operation to basically feed this information to however the model has been implemented.

But my point here when I use the word "off-the-shelf," it is a trivial operation and one that any person of ordinary skill in the art would be able to do. All the information is contained in the BAM file, and it can just be fed into the model.

There are high-level libraries that enable you to get all the information from a BAM file, and you put it in the model. This is really a very trivial step.

EX1002, ¶82; EX2037, 102:15-104:8; Pet., 29; Reply, 3-4

Prior Art Confirms Inputting BAM Data to Model Trivial

kineticsTools:

Example Usage

Basic use with BAM input, GFF+HDFS output:

```
ipdSummary aligned.bam --reference ref.fasta --identify m6A,m4C --gff basemods.gff --csv_h5 kinetics.h5
```

With cmp.h5 input, methyl fraction calculation and GFF+CSV output:

```
ipdSummary aligned.cmp.h5 --reference ref.fasta --identify m6A,m4C --methylFraction --gff basemods.gff --csv kinetics
```

Inputs

Aligned Reads

A standard PacBio alignment file - either AlignmentSet XML, BAM, or cmp.h5 - containing alignments and IPD information supplies the kinetic data used to perform modification detection. The standard cmp.h5 file of a SMRTportal jobs is data/aligned_read.cmp.h5.

Wenger and Deep Variant:

poor for indels. When DeepVariant is trained on CCS reads, performance improves dramatically. As more CCS datasets are made available, both model-based callers like GATK and learning-based callers like DeepVariant will have the opportunity to improve on the performance reported here, including the incorporation of haplotype phase information and evaluating and training against updated GIAB benchmarks that correct errors using CCS reads. Further, advances in sequencing chemistry or consensus base calling (such as the application of deep learning) that reduce the residual indel errors in CCS reads also could improve variant calling performance.

DeepVariant is a deep learning-based variant caller that takes aligned reads (in BAM or CRAM format), produces pileup image tensors from them, classifies each tensor using a convolutional neural network, and finally reports the results in a standard VCF or gVCF file.

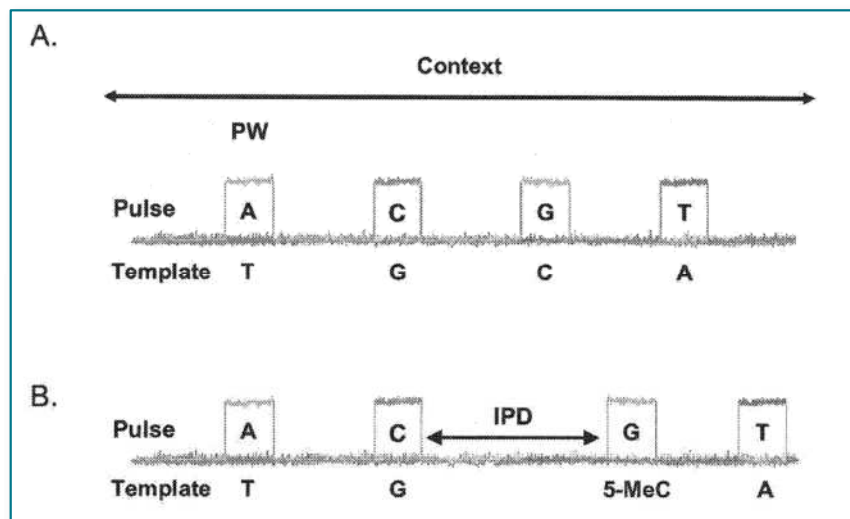
DeepVariant supports germline variant-calling in diploid organisms.

- NGS (Illumina) data for either a whole genome or whole exome.
- RNA-seq Case Study for Illumina RNA-seq.
- PacBio HiFi data, see the PacBio case study.

EX1021, 2; EX1032, 2; EX1036, 1; EX1002, ¶45; Reply, 4; Pet., 10-11

No Teaching Away from Nucleotide Identity

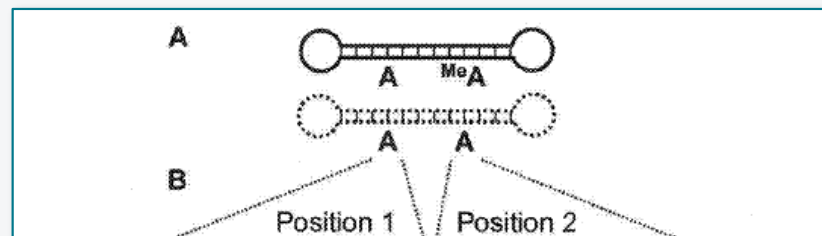
- Flusberg Publications taught to use nucleotide identity



specific embodiments. Different types of reaction data can be combined to detect various kinds of modifications, and in certain embodiments more than one type of modification can be detected and identified during a single reaction on a single template. Such variations to the detailed embodiments of the invention will be clear to one of ordinary skill based upon the teachings provided herein.

In certain embodiments, redundant sequence information is generated and analyzed to detect one or more modifications in a template nucleic acid. Redundancy can be achieved in various ways, including carrying out multiple sequencing reactions using the same original template, e.g., in an array format, e.g., a ZMW array. In some embodiments in which a lesion is unlikely to occur in all the copies of a given template, reaction data (e.g., sequence reads, kinetics, signal characteristics, signal context, and/or results from further statistical analyses) generated for the multiple reactions can be combined and subjected to statistical analysis to determine a

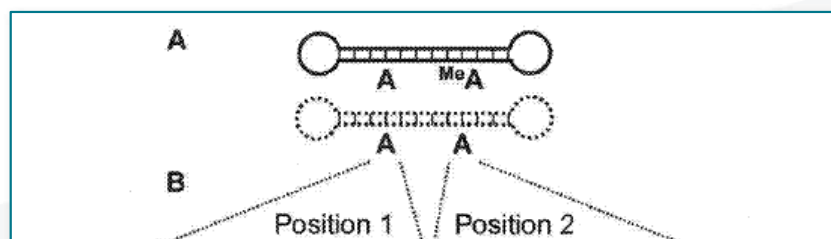
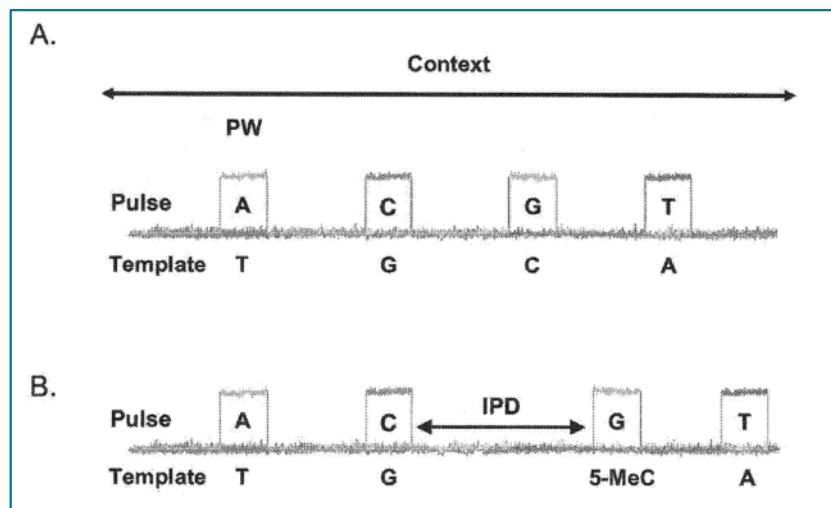
polymerase active site²⁹. Consistent with this model, the kinetic impact of methylation is not restricted to the nucleotide incorporation opposite the modified base. Additionally, the IPD ratio patterns differed for the two methylated positions. As the only source of difference between these two loci in all three template types was the local sequence context (template sequences are listed in Supplementary Note 1), we conclude that, in general, the kinetic signatures of methylation will be sequence context-dependent.



Reply, 7-11; EX1006, 462-63; EX1007, 14:57-15:8; Figs. 2, 14

No Teaching Away from Nucleotide Identity

Flusberg-338:



Dr. Koo:

Q. How is sequence context described here?

A. "The term 'sequence context' can refer to the base compositions A, C, G, or T, and the base orders in a stretch of DNA. Such a stretch of DNA could be surrounding a base that is subjected to or the target of base modification analysis.

"For example, the sequence context can refer to bases upstream and/or downstream of a base that is subjected to base modification analysis."

Q. Okay. The two sequences here are being

compared based on the identity of the nucleotide in the sequencing, correct?

A. That's correct.

EX1007, Figs. 2, 14; EX1049, 43:5-13, 88:15-18

Dr. Koo's Opinions Based on Legal and Factual Error

- Dr. Koo disregarded disclosure because he thought it was not real data

- Cross, re Flusberg-338, Figs. 11 and 14:
- Flusberg-338:

For example, FIG. 14 provides actual data showing that for two different positions in a single circular template, one always unmethylated, and one differentially methylated, an increase in the number of reads for the template corresponds to an increased resolution between IPDs for methylated vs. unmethylated adenosines. If the underlying distributions are

distribution. FIG. 11 shows actual molecular consensus distributions for methylcytosine, given the underlying gamma-like weighted IPD distributions of individual reads, but in this figure only the means of these underlying distributions were utilized. The plotted distributions could become even more well-separated if other characteristics had been taken into account. The data used to generate FIGS. 14 and 11 is more fully described in the Examples herein.

Q. All right, and a visual alignment would inspect whether the A corresponds to an A, the T to a T, and so forth, right?

A. That would be the general idea.

Q. When you say, "not realistic," you mean it's not -- it's not experimental data?

A. This is just a toy schematic. Exactly. That shows that it's not experimental data. This is --

Q. Okay.

A. -- an idealized version of what data could potentially look like in -- in a -- you know, it -- basically with no basis --

Q. Okay.

A. -- because there's no real data to support it yet.

Q. Okay, but you're --

It may be going far to say it's not realistic?

You're saying it's not experimental data?

A. I would -- I would agree with that. I wouldn't -- I wouldn't disagree with that statement.

Q. Okay, and then the next sentence where we're reading states "Otherwise, the two templates are identical." Do you see that?

A. Yes. And, of course, because this is a cartoon schematic, one can't verify that. One just has to assume based on the depiction of the artist what they

EX1049, 97:18-99:11, 103:11-14; EX1007, 65:47-66:3

Prior Art Corroborates Including Nucleotide Identity

- kineticsTools:

Studies of the relationship between IPD and sequence context reveal that most of the variation in mean IPD across a genome can be predicted from a 12-base sequence context surrounding the active site of the DNA polymerase. The bounds of the relevant context window correspond to the window of DNA in contact with the polymerase, as seen in DNA/polymerase crystal structures. To simplify the process of finding DNA modifications with PacBio data, the tool includes a pre-trained lookup table mapping 12-mer DNA sequences to mean IPDs observed in C2 chemistry.

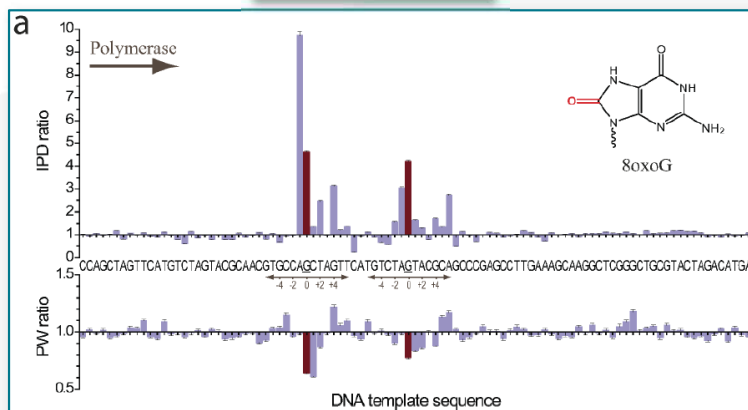
There are a few features of PacBio data that require special attention in order to achieve good modification detection performance. kineticsTools inspects the alignment between the observed bases and the reference sequence -- in order for an IPD measurement to be included in the analysis, the PacBio read sequence must match the reference sequence for k around

- White Paper:



The local sequence context of unmodified DNA influences polymerase dynamics. As such, the *in silico* control uses a predicted value of the mean IPD at each reference position that is based on the local sequence context.

- Clark:



pulses. The length of time that the polymerase retains a nucleotide bound in its active site (pulse width, PW), and the time interval between successive nucleotide-bound states (interpulse duration, IPD) are the principal pulse metrics used in the analysis to ascertain that the polymerase kinetics was altered in the modification-containing template when compared to an unmodified control template [5]. Because the DNA polymerase is in contact with the modified base over a region of ~11 bases [8], the kinetic effects are not necessarily restricted to the nucleotide incorporation opposite of the modified base. The magnitude and extent of the kinetic signature is dependent on the type of the base modification and the local sequence context. This results in distinct, more complex kinetic signatures that can be used to discriminate between different chemical base modifications [5].

EX1008, 1-2, Fig. 1; EX1016, 3; EX1021, 1-2; Reply, 12-13

PO Mischaracterizes Dr. Dessimoz's Testimony

- Sur-reply:

Petitioner's own expert recognized that kineticsTools includes sequence alignment [EX1002, ¶37], but testified that kineticsTools does not disclose including nucleotide identity as a property in an input data structure just like the Flusberg publications [EX2037, 193:11-13].

- Dr. Dessimoz:

THE WITNESS: So I'm a little bit uncomfortable with the language, this dichotomy of "fixed" versus "variable."

What I can say is that, as already taught in Flusberg, the sequence context is very important. Either you control it through some experiments, or if you can get this information from the base calling, then you certainly want to take that into account.

And it looks like this is an instance of the latter. But again, it's the same idea, which is to predict a modification, you have to take into account the sequence context, the pulse width, IPD distance, as taught in the Flusberg publication.

Sur-reply, 12; EX2037, 193:11-194:8

No Teaching Away from PW

- Each of Flusberg Publications and PacBio-BAM teaches to include PW for modification detection

Flusberg:

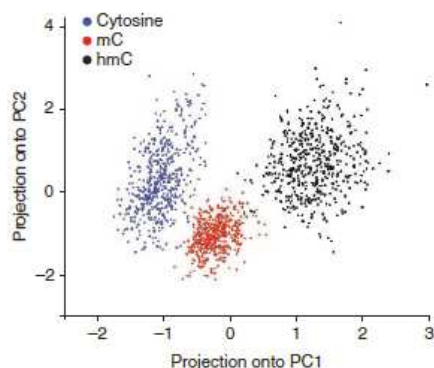


Figure 3 | Principal component analysis of cytosine, mC and hmC IPD and pulse width signatures. Each principal component is a linear combination of the mean IPD and pulse width at positions 71–79, which surround the variably modified template position 73. The weightings of IPD and pulse width at each position are shown in **Supplementary Table 1**. Data points on the plot were computed by projecting a random 20% subsample of the IPD and pulse width values onto the first two principal components (PC1 and PC2) and then converting to a z score.

PacBio-BAM:

Encoding of kinetics pulse features

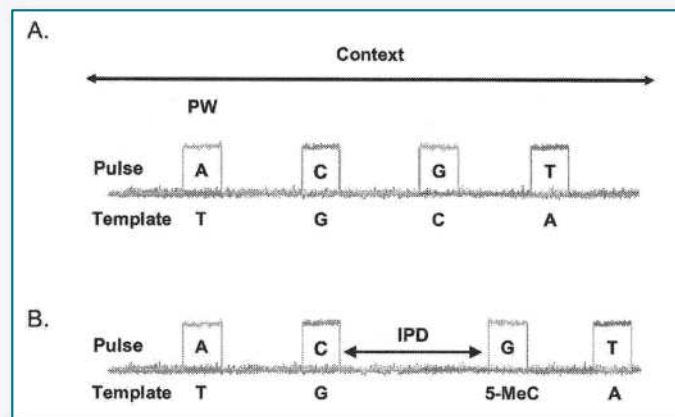
Interpulse duration (IPD) and pulsewidth are measured in frames; natively they are recorded as a uint16 per pulse/base event. They may be encoded in BAM read tags in one of two fashions:

- losslessly as an array of uint16; necessary for PacBio-internal applications but entails greater disk space usage.
- lossy 8-bit compression stored as a uint8 array, following the codec specified below ("codec V1"). Provides a substantial disk-space savings without affecting important production use cases (base modification detection).

Flusberg-338:

The two templates are identical except for the hydroxymethylcytosine in the hydroxymethylated template, which are indicated by arrowheads in FIG. 16. Templates comprising 5-hydroxymethylcytosine bases were also tested and the presence of these modifications was shown to have an effect on pulse width. FIG. 17 provides a plot of pulse width ratio (pulse width for methylated template divided by pulse width for unmethylated template) vs. template position where the modified positions comprise 5-hydroxymethylcytosine bases. Variably hydroxymethylated positions are indicated by the arrowheads.

tions, as well. Further, in light of the above teachings it will be clear to one of ordinary skill that the approach can be extended to pulse metrics other than IPD, such as pulse width, branch rate, mismatch rate, deletion rate, etc. In addition, the



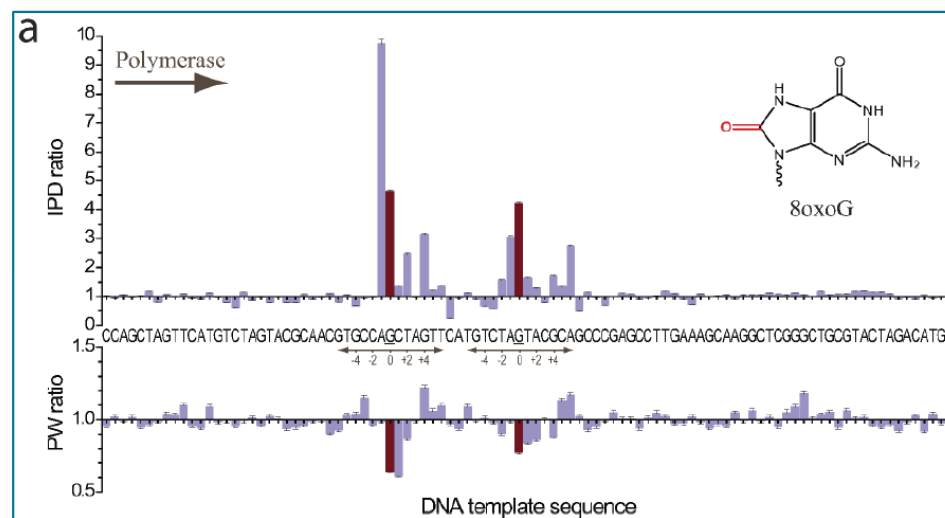
EX1006, Fig. 3; EX1007, 67:64-67, 78:7-12, Fig. 2; EX1011, 8; Reply, 14-15

No Teaching Away

Prior Art Corroborates Including PW

- Clark also teaches using PW in combination with IPD and sequence context

Recently, single-molecule, real-time (SMRT) DNA sequencing has been described for the direct detection of methylated and hydroxymethylated DNA bases [5]. In SMRT sequencing, the progression of single molecules of DNA polymerase is monitored in real time during base incorporations using fluorescent phospholinked nucleotides [6,7]. The dynamics of DNA polymerization is thereby recorded in the form of a train of fluorescent pulses. The length of time that the polymerase retains a nucleotide bound in its active site (pulse width, PW), and the time interval between successive nucleotide-bound states (interpulse duration, IPD) are the principal pulse metrics used in the analysis to ascertain that the polymerase kinetics was altered in the modification-containing template when compared to an unmodified control template [5]. Because the DNA polymerase is in contact with the modified base over a region of ~11 bases [8], the kinetic effects are not necessarily restricted to the nucleotide incorporation opposite of the modified base. The magnitude and extent of the kinetic signature is dependent on the type of the base modification and the local sequence context. This results in distinct, more complex kinetic signatures that can be used to discriminate between different chemical base modifications [5].



across the 8oxoG modification in the template. Additional characteristic kinetic signals surrounding the site of the base modification were observed, with a relatively strong IPD ratio increase for nucleotide incorporations just prior to the site of the modification (position -1), and smaller signals up to 6 bases past the modification site (positions +1 to +6). For both modification sites, a decrease in the IPD ratio for the +7 position was observed, i.e. the polymerase bound the nucleotide faster compared to the unmodified control. Pulse widths were also affected by the 8oxoG modifications albeit with smaller magnitude, with decreased PW values spanning positions zero to +2 compared to the control. In addi-

No Teaching Away

Prior Art Corroborates Including PW

- White Paper teaches using tools that evaluate PW, IPD, and sequence context in “typical experiment for detecting DNA base modifications”

using the *in silico* control. Future commercial products are expected to improve use of the sequence context to better identify and interpret modification signals.

Based on the measurement and analysis principle outlined above, a typical experiment for detecting DNA base modifications can be designed as follows:

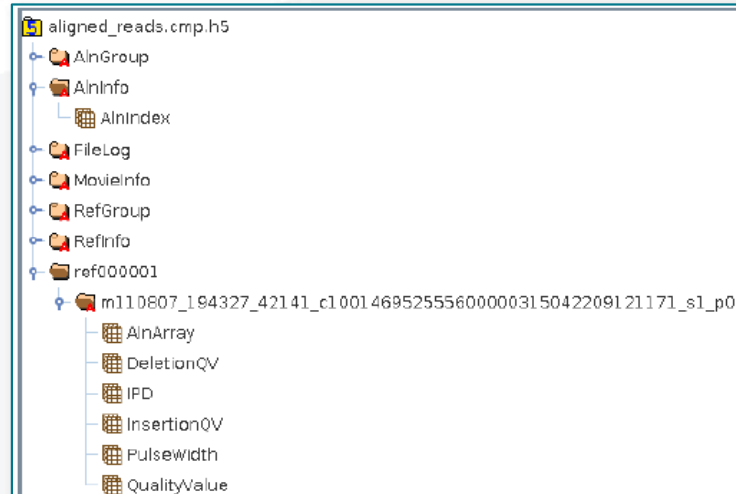
- Obtain native DNA of interest and prepare a SMRTbell™ library for SMRT sequencing (see the Pacific Biosciences Sample Preparation and Sequencing Guide).
- If using an amplified control design, use a small aliquot of the original DNA sample and perform a whole-genome amplification (WGA) reaction to obtain the control DNA sample lacking any base modifications. Then prepare a SMRTbell library for this sample.
- Perform adequate SMRT sequencing (for both samples when using an amplified control) to obtain the necessary coverage needed for the modifications under study, and sufficient overall coverage to characterize the genome (or portion of the genome) of interest.
- Use the bioinformatics tools to perform kinetic analysis for base modification⁹.

9. SMRT Analysis version 1.3.1 or higher, and/or “DNA Modification Detection with SMRT Sequencing using R” found at <https://github.com/PacificBiosciences/R-kinetics>.

2.3 Context-specific Effects

Finally, we investigate the effect of sequence context on IPD and pulse width distributions. Alignment-level data from a cmp.h5 file are always stored with respect to the bases being incorporated. Therefore, when one retrieves an alignment from the cmp.h5 file, if that alignment is labeled as a reverse strand alignment: `getTemplateStrand(cmpH5) == 1`, then the reference sequence is reverse complemented rather than the read. The importance of this representation is that we always store the data (e.g., alignments, IPDs, pulse widths, etc.) in the direction in which the bases are incorporated.

sequence. The mapping between the canonical “ref000001” and the name in the FASTA file can be found using the `refGroup` function. For our work, the most pertinent datasets are: `AlnArray`, `IPD` (interpulse duration), and `PulseWidth`.



EX1016, 4, 5; EX1047, 4-5, 11; Reply, 15-16

No Teaching Away

PO Mischaracterizes Schadt

- Schadt:

the interpulse duration or IPD). While there is a fair degree of stochastic variation in the IPD and PW values over different reads covering a common genomic region, reproducible variations in IPD and PW are induced as a result of modifications to the bases, as has been previously shown (Flusberg et al. 2010). Given the high degree of regulation involved in the base incorporation process, PW is not as sensitive to shifts in sequence context induced by modifications as IPD (Flusberg et al. 2010). Therefore, for the purposes of developing an initial model of kinetic variation, we restrict attention to IPDs.

- Dr. Dessimoz:

38. Other references further illustrate models which receive as input a combination of kinetic parameters (including IPD and PW) and the local sequence context over a window of nucleotides. For example, Schadt teaches that both PW and IPD are affected by modifications, with IPD being more strongly affected, and further teaches that polymerase kinetics “is exquisitely sensitive to local sequence.” EX1010, 130 (citing Flusberg, EX1006). Indeed, Schadt teaches that the nucleotide identities over a 10-nucleotide window including a nucleotide of interest produces “>80% of the variance in the IPD measures.” *Id.*; see also

EX1002, ¶138; EX1010, 130; Reply, 15-16

No Objective Indicia

- POR:

A. The Method of Claim 1 Solved the Long-Felt Unsolved Need for Direct Detection of DNA Modifications by SMRT Sequencing Without Bisulfite Conversion.....37

- FALSE

B. PacBio Tried and Failed to Directly Detect DNA Methylation Without Bisulfite Conversion Through SMRT Sequencing.....49

- FALSE

C. The Holistic Kinetic Model Was Praised by PacBio and others in the field.....50

- FALSE

D. PacBio Copied the Holistic Kinetic Model Disclosed in the Tse Paper.....54

- FALSE

E. The Method of Claim 1 Had Unexpected Results Compared to Flusberg 62

- FALSE

POR, i-ii; Reply, 16-27

Objective Indicia: No Nexus

- Dr. Koo performed no nexus analysis:
- Dr. Koo admits Tse is at best an embodiment of claim 1
 - EX2002:

VII. Tse et al. discloses an embodiment of claim 1. 31

Q. Must the model of Claim 1 be a convolutional neural network?

A. From my understanding, it does not have to be limited to a convolutional neural network model.

Q. Looking through this discussion, do you have a discussion of the legal concept of nexus here?

A. I'll have to doublecheck.

Q. Can you do that for me, please?

(Pause.)

THE WITNESS: In my reading right now, I do not see the word "nexus."

BY MR. ROSATO:

Q. There's no discussion of the legal standard for "nexus," correct?

MR. HAMILTON: Objection, vague. Calls for a legal conclusion.

THE WITNESS: I do not a word that says "nexus."

Q. Is the window of Claim 1 limited to a 21-nucleotide window?

A. Claim 1 doesn't specify a particular window size.

EX1049, 73:4-7, 78:9-12, 149:9-22; EX2002, i; Reply, 16

Objective Indicia: Not Commensurate with Claims

- POR: same drawbacks noted above in addition to fragmenting the DNA.”). In sum, at the time of the '794 patent, the long-felt need for direct detection of DNA methylation, without bisulfite conversion, through SMRT sequencing and not been satisfied by others.

- Claims do not exclude bisulfite conversion
- Claims not limited to methylation, 5mC, SMRT sequencing, or even DNA

Q. Does Claim 1 recite the term "SMRT sequencing"?

A. I don't recall it explicitly stating "SMRT sequencing," which -- SMRT sequencing meaning PacBio's commercial product?

A. It says "A modification of a nucleotide in a nucleic acid molecule."

Q. Right, so we discussed this earlier. DNA is a nucleic acid molecule, right?

A. DNA is a nucleic acid molecule.

Q. And RNA is a nucleic acid molecule, correct?

A. RNA is as well.

Q. Is "bisulfite sequencing" mentioned anywhere in Claim 1?

A. I don't see "bisulfite sequencing" mentioned in Claim 1.

Q. And we identified 14 different examples of methylation modifications that were listed in the '794 patent. Do you remember that?

A. There were 14 different types of methylation listed.

Q. Yeah, and another type of modification that was discussed is an oxidation. Do you recall that discussion?

A. I recall that we -- we -- we discussed that.

Q. Okay, and the claim limitation we're looking at here states "Modification," correct?

A. Yes, the modification is stated here.

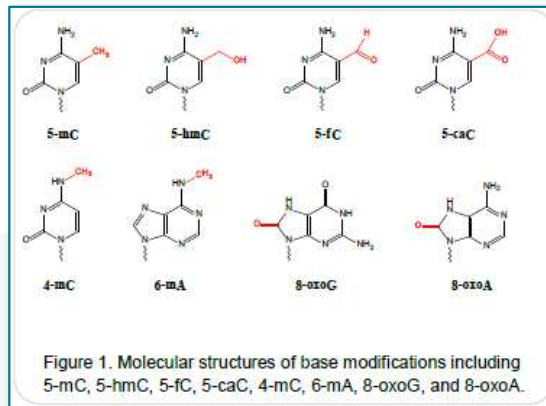
Objective Indicia: Alleged Need Already Met

- Flusberg:

We describe the direct detection of DNA methylation, without bisulfite conversion, through single-molecule, real-time (SMRT) sequencing. In SMRT sequencing, DNA polymerases catalyze

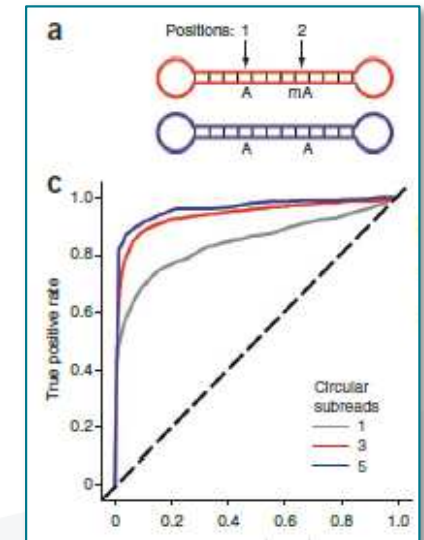
- White Paper:

samples². SMRT sequencing, in contrast, does not require base conversion within the source material in order to detect base modifications. Instead, the



- Clark:

Recently, single-molecule, real-time (SMRT) DNA sequencing has been described for the direct detection of methylated and hydroxymethylated DNA bases [5]. In



Objective Indicia: No Failure

PO Mischaracterizes Schadt

- POR:

But, the Flusberg method failed to solve the need for the direct detection of DNA methylation, without bisulfite conversion, through SMRT sequencing because, as explained by Flusberg's co-authors in Schadt three years later "[p]resently, only 5-mC can be routinely detected using NGS technology and the only after the samples have been bisulfite treated, a procedure that creates a number of experimental and informatics challenges." EX1010 at 129; EX2041.

- Schadt:

DNA sequencing carried out using current first and second (or next) generation sequencing (NGS) instruments has achieved significant success in providing DNA sequences in which the A's, G's, C's, and T's comprising any given DNA template of interest are very consistently called at accuracies that exceed 99%, enabling an

- Dr. Koo:

Q. Okay, so am I clear in understanding that there's a difference between second-generation sequencing and third-generation sequencing?

A. And -- and what was the question?

Q. So the first paragraph in Schadt discusses second-generation sequencing, correct?

(Pause.)

THE WITNESS: I -- I just need a second to read over this to make sure that's exactly what they talk about.

It's -- it's mostly about, yeah, second-generation -- first- or second-generation sequencing.

Presently, only 5-mC can be routinely detected using NGS technology and then only after samples have been bisulfite treated, a procedure that creates a number of experimental and informatics challenges (Cokus et al. 2008). Thus, new technologies and assays are needed to fully characterize DNA and RNA sequences if we hope to completely characterize variation in genomes and transcriptomes.

Single-molecule real-time (SMRT) technology is the first commercially available platform to directly observe single molecules of DNA polymerase as they synthesize DNA sequences, not

POR, 42; EX1049, 56:3-15; EX1010, 129; Reply, 20

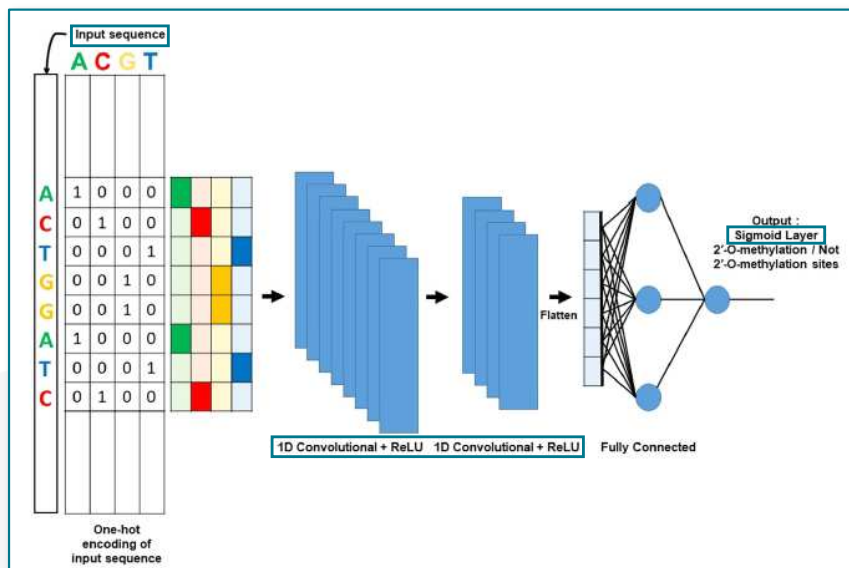
Objective Indicia: No Copying

- Admitted differences in window size, elements, data structure, and CNN:

PO curiously alleges copying of Tse's HK model by Dr. Wenger but readily admits he used a different data structure (e.g., 16 versus 21 nucleotide window, different matrix organization) and a different CNN ("TensorFlow CNN"). EX2041, ¶¶81-83, 85.

- Alleged similarities -> conventional structures of prior art CNNs:
 - input layer
 - convolutional layers
 - sigmoid function

Tahir:



Dr. Koo:

Q. Okay. So typically, a convolutional neural network has a plurality of layers; is that accurate?

A. I would agree that typically, a convolutional neural network would have multiple convolutional layers.

Q. And does a convolutional neural network typically have an input layer?

A. That's correct.

Q. Does it typically have an output layer?

A. Typically, yes, if trained in a supervised manner.

EX1049, 80:18-824; EX1035, Fig. 1; Reply, 21-24

Objective Indicia: No Praise

- PO alleges praise from Mr. Henry in 2020 earnings call
- HiFi sequencing product released in 2019
- Praise directed to PacBio's own products:
 - Mr. Henry described “the power and the potential of PacBio HiFi sequencing technology”
 - Mr. Henry shared “customer highlights”
 - Customers using PacBio's product included Dr. Lo and HudsonAlpha Institute of Biology

Dr. Koo:

Q. Okay. Are you familiar with PacBio's HiFi sequencing product?

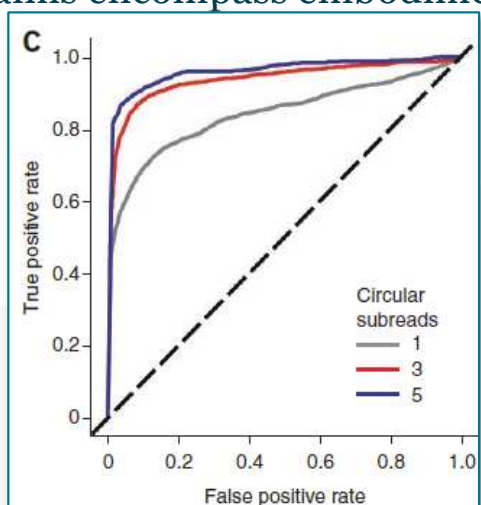
A. I'm -- I'm not familiar.

Reply, 24-25; EX1049, 36:4-6

Objective Indicia: No Unexpected Results

- Comparison in Fig. 29 disconnected from and not commensurate with claims
 - Comparison between neural network and PCA, but claims encompass both
 - Comparison does not reflect Flusberg Publications' actual AUC
 - Flusberg-338 not limited to PCA, but expressly discloses using neural network
 - Flusberg reports AUC of 0.96, superior to best performance reported in '794 patent
 - Patent figures show model type affects performance (e.g., logistic regression lowers AUC to 0.83)
 - Claims encompass embodiments with worse AUC (AUC=0.5 for many window sizes)

Flusberg:



(Fig. 4c) in a particular context. The normalized area under the ROC curve was 0.80 after the first circular subread but increased to 0.92 and 0.96 after three and five circular subreads, respectively.

'794 patent:

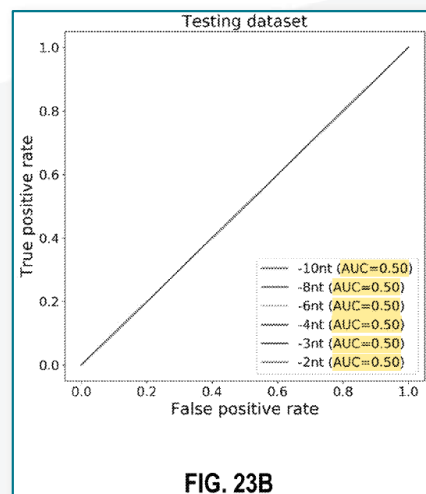


FIG. 23B

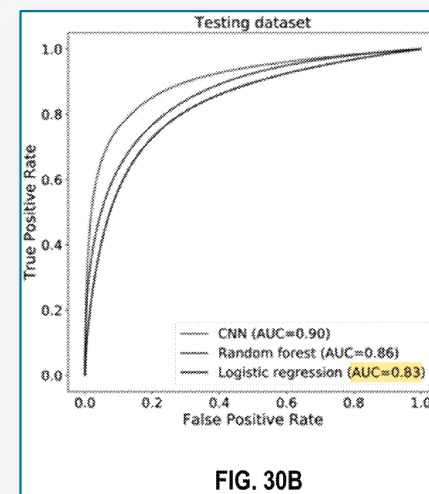


FIG. 30B

Reply, 25-27; EX1001, Figs 23B, 30B; EX1006, 463-64, Fig. 4c