

Filed on behalf of: The Chinese University of Hong Kong

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

PACIFIC BIOSCIENCES OF CALIFORNIA, INC.

Petitioner,

v.

THE CHINESE UNIVERSITY OF HONG KONG

Patent Owner

IPR2024-00028
Patent 11,091,794

EXHIBIT 2002: DECLARATION OF DR. PETER K. KOO

TABLE OF CONTENTS

I.	Introduction	1
II.	Qualifications	2
III.	Bases of My Opinion	3
	A. Materials Considered.....	3
	B. Relevant Legal Principles.....	4
IV.	Retention and Compensation	6
V.	Person of Ordinary Skill in the Art.....	7
VI.	The '794 Patent.....	7
	A. Background	7
	B. Construction of claim 1	15
	C. The '794 patent discloses probing the predictive power of features.	24
	D. Prosecution history of claim 1	30
VII.	Tse et al. discloses an embodiment of claim 1.	31
VIII.	Flusberg does not teach or suggest to a person of ordinary skill in the art the “input data structure” of claim 1.	34
IX.	The '338 patent does not teach or suggest to a person of ordinary skill in the art the “input data structure” of claim 1.	47
X.	The PacBio BAM does not teach or suggest to a person of ordinary skill in the art the “input data structure” of claim 1.	73
XI.	A person of ordinary skill in the art would not have been motivated to combine Flusberg, the '338 patent, and the PacBio BAM to arrive at the invention of claim 1.....	76
XII.	Modification detection using the claimed input data structure provided unexpected results compared to Flusberg and the '338 patent.....	82

I, Peter K. Koo, declare as follows:

I. Introduction

1. I understand that Pacific Biosciences of California, Inc. (“PacBio”) has filed a petition of *inter partes* review before the Patent Trial and Appeal Board of the United States Patent and Trademark Office asserting that the claims of the U.S. Patent No. 11,091,794 (“the ’794 patent”) are unpatentable as obvious under 35 U.S.C. § 103 in light of Flusberg, et al., “*Direct Detection of DNA Methylation During Single-Molecule, Real-Time Sequencing*,” Nature Methods 7(6) (2010) (“Flusberg”) EX1006, U.S. Patent No. 9,175,338, Flusberg et al. “*Methods for Identifying Nucleic Acid Modifications*” (“the ’338 patent”) EX1007, and BAM Format Specification for PacBio, PacBioFileFormats 5.0 Documentation, <http://web.archive.org/web/20170916112547/http://pacbiofileformats.readthedocs.io/en/5.0/BAM.html> (archived September 16, 2017) (“the PacBio BAM”) EX1011. I have been asked by counsel for the Chinese University of Hong Kong (“CUHK”) to express my opinions on the patentability of the claims of the ’794 patent.

2. This Declaration is based upon the information currently available to me. To the extent that additional information becomes available, I reserve the right to continue my investigation and study.

3. I have been asked to focus my opinion of Ground 1 of the petition, namely that PacBio asserts that claims 1-2, 7, 10-12, 14-15, and 17-19 of the ’794

patent are unpatentable under 35 U.S.C. § 103 as obvious based upon Flusberg, the '338 patent and the PacBio BAM. In particular, I have been asked to focus on claim 1 of the '794 patent with respect to Ground 1.

4. In my opinion, claim 1 of the '794 patent is non-obvious over Flusberg, the '338 patent, and the PacBio BAM. My opinion is the same if Flusberg or the '338 patent is considered alone or together with the PacBio BAM.

II. Qualifications

5. I am an Assistant Professor, at the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory, Cold Spring Harbor, New York. I am also a member of the Cold Spring Harbor Laboratory Cancer Center. I am an affiliated Faculty at the Department of Applied Math and Statistics, the Laufer Center for Physical and Quantitative Biology, and the Graduate Program in Genetics at Stony Brook University, Stony Brook, New York.

6. I received my Ph.D. in Physics from Yale University in 2015 and a Master of Science degree in Physics from Yale University in 2008. I received an undergraduate degree in Physics from the University of California at Berkeley in 2006. I also received a Master of Financial Engineering degree from the Haas School of Business at the University of California at Berkeley in 2010. From 2015-2019, I was a Postdoctoral Research Associate in the laboratory of Dr. Sean R. Eddy at the Howard Hughes Medical Institute and Harvard University. My postdoctoral

research focused on the development of computational methods for RNA, protein, and DNA sequence analysis.

7. Since 2019, I have headed up a laboratory at the Simons Center for Quantitative Biology at Cold Springs Harbor Laboratory. My laboratory studies gene regulation through a computational lens using data-driven machine learning solutions. Our approach develops deep learning methods to infer sequence-function relationships that underlie high-throughput functional genomics data. Through model interpretability, we aim to elucidate cis-regulatory mechanisms -- the complex coordination of sequence elements such as motifs -- with a broader aim of advancing precision medicine for complex diseases, including cancer. My laboratory is part of the National Cancer Institute designated Cancer Center at Cold Spring Harbor Laboratory.

8. Exhibit 2003 is a copy of my *curriculum vitae* which includes further details of my education, training, experiences, publications and additional qualifications.

III. Bases of My Opinion

A. Materials Considered

9. A list of the materials I considered, in addition to my experience, education, and training, in providing the opinions set forth in this declaration is attached as Exhibit 2032.

B. Relevant Legal Principles

10. I am not an attorney, but in preparing and forming my opinions, I have been informed of certain legal principles. I have applied my understanding of those principles and taken them into account when forming the opinions described herein. My understanding of the relevant legal principles is summarized below.

11. I understand that claim terms are generally given their ordinary and customary meaning as understood by a person of ordinary skill in the art at the time of the invention when read in the context of the specification and prosecution history unless a patentee sets out a different definition or clearly disavows claim scope.

12. I further understand that extrinsic evidence such as expert and inventor testimony, dictionaries, and learned treatises can help determine the meaning of claim terms, although that evidence is less significant than the claims, specification, and prosecution history.

13. I have been informed that PacBio bears the burden of proving unpatentability by a preponderance of evidence. I have been told that this means that PacBio must prove that more likely than not that claim 1 of the '794 patent is obvious in light of Flusberg, the '338 patent and the PacBio BAM.

14. I understand that my opinions regarding patentability are from the viewpoint of a person having ordinary skill in the field of the technology of the patent

as of the time of the invention. For the purposes of this declaration, I have assumed that date is the earliest priority date of the '794 patent which is August 16, 2019.

15. I understand that for a patent claim to be anticipated, each claim limitation must be present in a single piece of prior art. It is my understanding from counsel that a prior art reference anticipates a claimed invention if the prior art reference disclosed each of the claimed limitations of the invention either expressly or inherently.

16. I understand that if the differences between the claimed invention and the prior art are such that the claimed invention as a whole would have been obvious before the effective filing date of the claimed invention to a person having ordinary skill in the art to which the claimed invention pertains, the claimed invention is obvious.

17. I understand there are four fact-based inquiries involved in determining patent obviousness. These include: (1) the scope and content of the prior art; (2) the differences between the claimed invention and the prior art; (3) the level of ordinary skill in the art; and (4) objective indicia of non-obviousness. I have been informed that examples of objective indicia include unexpected results, commercial success of the invention, whether the invention satisfied a long felt need in the industry, failure of others to find a solution to the problem at hand, commercial acquiescence

via licensing, professional approval, unexpected results, and copying and praise by infringers.

18. I understand that even if all limitations of a claimed invention are disclosed by the prior art combination, the patent challenger must demonstrate an apparent reason to combine the known elements in the fashion of the patent claim at issue and that a person of ordinary skill in the art would have reasonable expectation of success in pursuing that combination.

19. I understand that a prior art reference teaches away from a modification of a prior art reference when a person of ordinary skill in the art would be discouraged from following the path set out in the reference, or would be led in a direction divergent from the path followed by and claimed in the patent.

IV. Retention and Compensation

20. I have been retained by counsel for the Chinese University of Hong Kong (CUHK) as a technical expert in this matter to provide my opinions relating to the '794 patent. I receive \$500 (USD) per hour for my services. My compensation is in no way dependent upon my opinions or the outcome of this case. I have no affiliations with CUHK, or its licensee Take2 Technologies Limited, any affiliates presently known to me, or the named inventors on the '794 patent.

V. Person of Ordinary Skill in the Art

21. As stated above, I have been informed by counsel that the obviousness analysis is to be conducted from the perspective of a person of ordinary skill in the art (a “person of ordinary skill,” “skilled artisan,” or “POSA”) at the time of the invention. I have also been informed by counsel that in defining a person of ordinary skill the following factors may be considered: (1) the educational level of the inventor; (2) the type of problems encountered in the art; (3) prior art solutions to those problems; (4) rapidity with which innovations are made; and (5) sophistication of the technology and educational level of active workers in the field.

22. In my opinion, a skilled artisan would have had the qualifications set forth in Dr. Dessimoz’s declaration at paragraph 75. EX1002, pp. 40-41, ¶75.

VI. The ’794 Patent

A. Background

23. The ’794 patent is titled “*Determination of Base Modifications of Nucleic Acids*” and was filed as U.S. Patent Application No. 16/995,607 on August 17, 2020. The ’794 patent claims priority of several provisional applications filed between August 16, 2019 and July 13, 2020.

24. The ’794 patent issued with 19 claims. Claim 1 is the sole independent claim which reads as follows:

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 identify 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 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 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.

25. Claim 1 of the '794 patent recites a novel set of features which had not been used before in combination to detect nucleotide modifications with a model. Specifically, those features are structured in an input data structure which includes a window and the properties of each nucleotide in the window, namely, the identity of the nucleotide, the 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. For simplicity, I will use the following shorthand terms: pulse width or PW for the width of the pulse, and IPD for interpulse duration.

26. The selection of features to combine in a data structure is a fundamental problem in model training. Because some input data, also referred to as input features, carry redundant information and some do not have any predictive information, these features can introduce excess noise, the selection of the appropriate features as data inputs for a trainable model, which is also referred to as feature engineering, feature construction, feature selection, or feature extraction, is unpredictable. In order to select appropriate features, the predictive power of the features needs to be assessed in combination with the training data, model choice, and the task. As provided in “Feature Engineering for Machine Learning” by Zhang et al., “[t]he number of features is also important. If there are not enough informative features, then the model will be unable to perform the ultimate task. If there are too many features, or if most of them are irrelevant, then the model will be more expensive and tricky to train. Something might go awry in the training process that impacts the model’s performance.” EX2006, p. 15.

27. Liu et al. probed Oxford Nanopore sequencing features, and found: “Our results show that base quality, deletion frequency and mismatch frequency alone were able to accurately predict the modification status with reasonable accuracy (70-86% accuracy, depending on the feature used) (**Figure 2D**, *see also*

Table SI and Methods). By contrast, we find that the current mean intensity values and current intensity standard deviation were poor predictors of the modification status of the k-mer (43-65% accuracy)." EX1038, pp. 3-4. Liu's findings illustrate the unpredictability of selecting the right combination of input data features.

28. Some of the challenges that researchers faced in selecting the appropriate features from the PacBio SMRT sequencing data set to train a model to detect base modifications are summarized in the '794 patent. EX1001, 17:7-19:50.

29. Because PacBio's petition is premised on Flusberg. I address the '794 patent disclosure directed to Flusberg.

30. At 22:31-38, the '794 patent (EX1001) discloses that:

Flusberg et al. 2010 did not use sequence context and IPD as independent inputs to detect a modification which resulted in a model that lacked practically meaningful accuracy for detection. Flusberg et al. only used sequence context to confirm 6 mA occurred in GATC. Flusberg et al. is silent as to using sequence context in combination with IPD as inputs to detect methylation status. EX1001, 22:31-38.

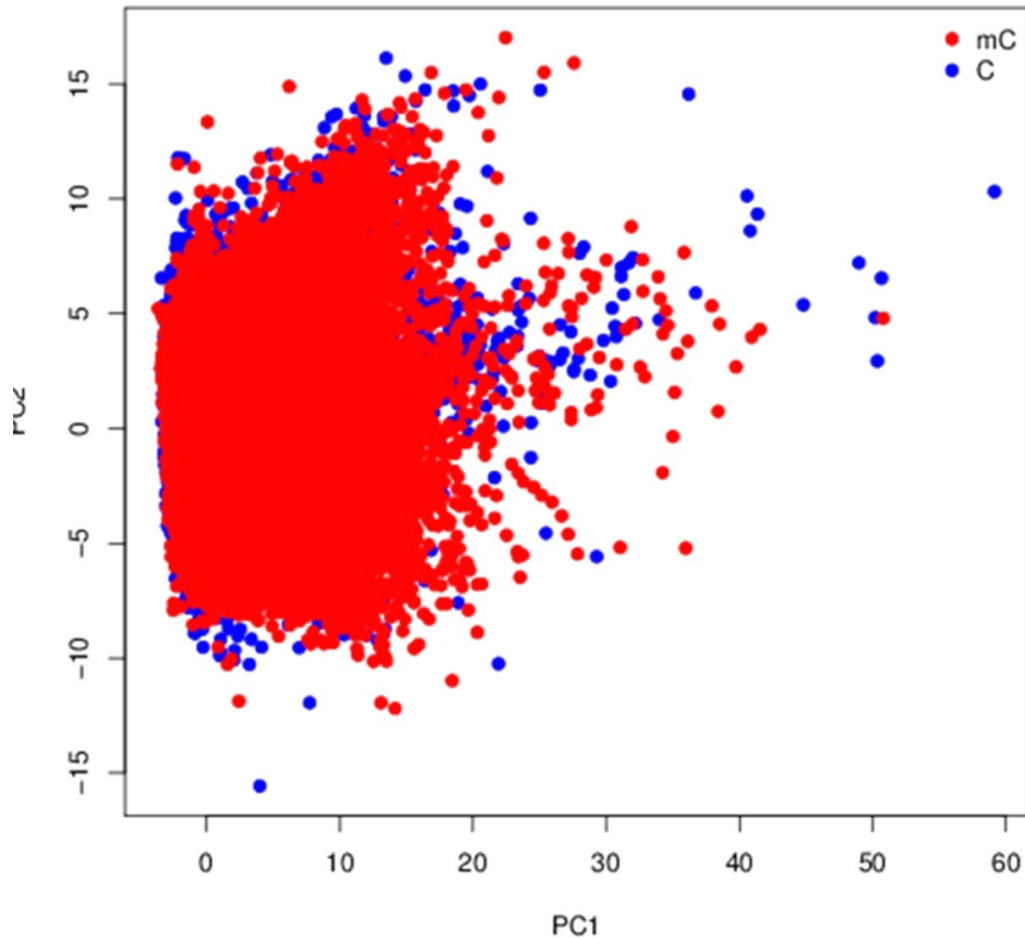
31. I agree with that statement. Flusberg did not use sequence context and IPD as independent inputs or features to detect a base modification. It is important to note that sequence context in Flusberg and the '794 patent are used differently. In the '794 patent, "[s]equence context refers to the base compositions (A, C, G, or T) and the base orders in a stretch of DNA." EX1001, 19:60-62. In other words,

sequence context refers to the identity of each nucleotide in order in a string of DNA. Claim 1 uses the nucleotide identity of each nucleotide in a defined window around a target nucleotide as a data input. Flusberg does not use nucleotide identity as an input for DNA modification detection. As I explain below, Flusberg typically uses sequence context to refer to a window of bases surrounding a methylation site, not as an input for DNA modification detection. The '338 patent uses sequence context in the same way as Flusberg, as I explain below.

32. The '794 patent tested the method in Flusberg and showed that principal component analysis using IPD and PW cannot differentiate 5-methylcytosine from unmethylated cytosine.

We also tested the method present in Flusberg et al. study (Flusberg et al., 2010). We analyzed a total of 5,948,348 DNA segments which were 2-nt upstream and 6-nt down-stream of a cytosine that was subjected to methylation analysis. There were 2,828,848 segments that were methylated and 3,119,500 segments that were unmethylated. As shown in FIG. 28, the signals deduced from the principal component analysis with the use of IPDs and PWs were found to be largely overlapping between fragments with methylated cytosines (mC) and unmethylated cytosines (C), suggesting that the method as described by Flusberg et al[.] lacks practically meaningful accuracy. These results suggested that the principal component analysis, which linearly combined PW and IPD values at the bases and neighboring bases, as used in Flusberg et al[.]’s study (Flusberg et al. 2010) could not reliably

or meaningfully differentiate 5-methylcytosine and unmethylated cytosines. EX1001, 35:21-37.

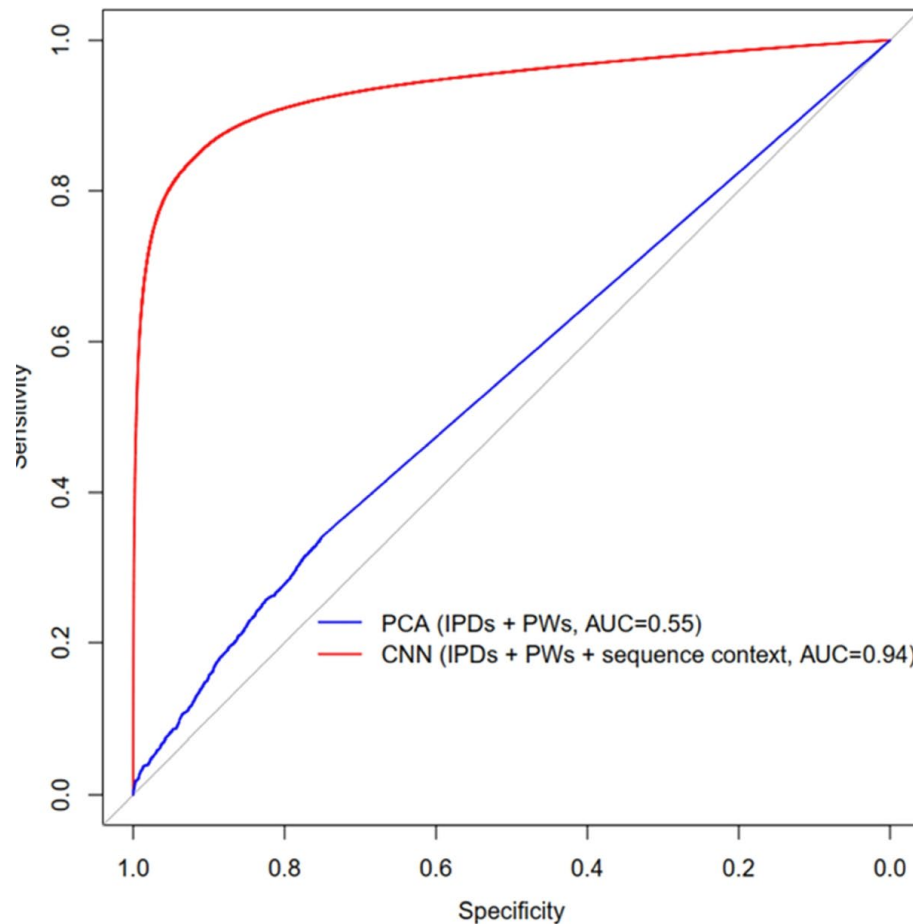


EX2001, FIG. 28.

33. As shown above in FIG. 28 from the '794 patent the principal component analysis of Flusberg was unable to meaningfully separate 5-methylcytosine from unmethylated cytosine.

34. The '794 patent also discloses that:

FIG. 29 shows that the AUC of the method based on the principal component analysis for which two principal components were used in Flusberg et al[.]’s study (Flusberg et al., 2010) involving IPDs and PWs were much less accurate (AUC 0.55) than the approach based on a convolutional neural network involving IPD and PW as well as sequence context as shown in our disclosure (AUC: 0.94). EX1001, 35:38-44.



EX2001, FIG. 29.

35. As shown above in FIG. 29, the method of claim 1 resulted in an AUC of 0.94 which a person of ordinary skill in the art would recognize as having a very

strong predictive power. In contrast, the principal component analysis method of Flusberg resulted in an AUC of 0.55, which is essentially random guessing. The reason the method of claim 1 has such a high AUC is because the inventors discovered, by systematically probing the predictive power of features, an optimal feature set to construct an input data structure that could accurately detect base modification. Those features are a window, and the properties of each nucleotide in the window, namely, the nucleotide identity, the position of the nucleotide with respect to a target position within the window, the PW, and the IPD.

B. Construction of claim 1

36. I understand that to properly construe a claim the words of the claim must be considered. And, in my review of claim 1, I use ordinary meaning of the terms as would be understood by a person of ordinary skill in the art as of August 16, 2019. That meaning remains the same today. Claim 1 recites “an input data structure.” The input data structure includes “a window of the nucleotides sequenced in [a] sample nucleic acid molecule.” Claim 1 further requires that the input data structure include certain properties including “the identity of the nucleotide” that was sequenced, the “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.” Thus, claim 1 recites an input data structure that includes a

number of nucleotides identified in order relative to a target base and the pulse width and the interpulse duration for each nucleotide.

37. Claim 1 continues and requires that the input data structure be input into a model. The model is trained by receiving multiple data structures as required by claim 1, “the model trained by: receiving a first plurality of first data structures.” Those first data structures have actual values for the nucleotide properties for defined input data structure (“each first data structure comprising values for the same properties as the input data structure”). Those values are derived from the data received from the sequencer – “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.”

38. According to the plain language of claim 1, the input data structure must be structured so that it can train a model. A person of ordinary skill in the art would understand that to train the model to detect modifications from an input data structure, the training data must be provided in the same format, *i.e.*, the same window with the same properties (nucleotide identity, nucleotide position relative to a target nucleotide position, nucleotide PW and IPD) as the input data structure that

would be used to detect DNA modifications (testing data). I note that Dr. Dessimoz agrees that the input data structure must be so structured.

As I explained for element 1.3, a person of ordinary skill in the art would have understood that to train a model to detect modifications from an input data structure, training data should be provided in the same format—and containing the same properties—as the input data structure that would be useful for methylation detection. Thus, the data structures used to train a model that would operate on the data structures I discussed above for element 1.2 would be trained on a data set composed of data structures containing the same properties – i.e., nucleotide identity, relative position to a target nucleotide, PW and IPD for each nucleotide in a window. EX1002, p. 79, ¶130.

39. And, after the model has been trained by the first data structures, data is received (nucleotide identity, nucleotide position in a nucleic acid, nucleotide PW and IPD) from a sequencer, the input data structure (window, nucleotide identity, nucleotide position relative to target, nucleotide PW and IPD) is created from the received data, the input data structure is input to a model, then the final step of the method is determining using the model, whether the modification is present in the nucleotide at the target position within the window in the input data structure.

40. The '794 patent discloses how to structure the input data structure to train a model.

In FIG. 4, the 3 bases from each side of a base that was subjected to base modification analysis would be defined as a measurement window **400**. In one embodiment, sequence context, IPDs, and PWs for these 7 bases (i.e., 3-nucleotide (nt) upstream and downstream sequence[s] and one nucleotide for base modification analysis) were compiled into a 2-dimensional (i.e., 2-D) matrix as a measurement window. In the example shown, the measurement window **400** is for one subread of the Watson strand. EX1001, 23:18-26.

41. The portion of FIG. 4 that illustrates the data matrix is shown below.

Measurement window 400

	G		A		T		N		A		C		T		402
	-3		-2		-1		0		+1		+2		+3		404
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	N
A	0	0	1.0	0.3	0	0	0	0	1.1	0.8	0	0	0	0	408
C	0	0	0	0	0	0	0	0	0	0	2.3	0.8	0	0	412
G	1.5	1.7	0	0	0	0	0.9	1.2	0	0	0	0	0	0	416
T	0	0	0	0	1.2	0.2	0	0	0	0	0	0	1.5	1.7	420

EX2001, FIG. 4 (partial).

42. The '794 patent explains the contents of the matrix:

The first row **402** of the matrix indicated the sequence that was studied. In the second row **404** of the matrix, the position 0 represented the base

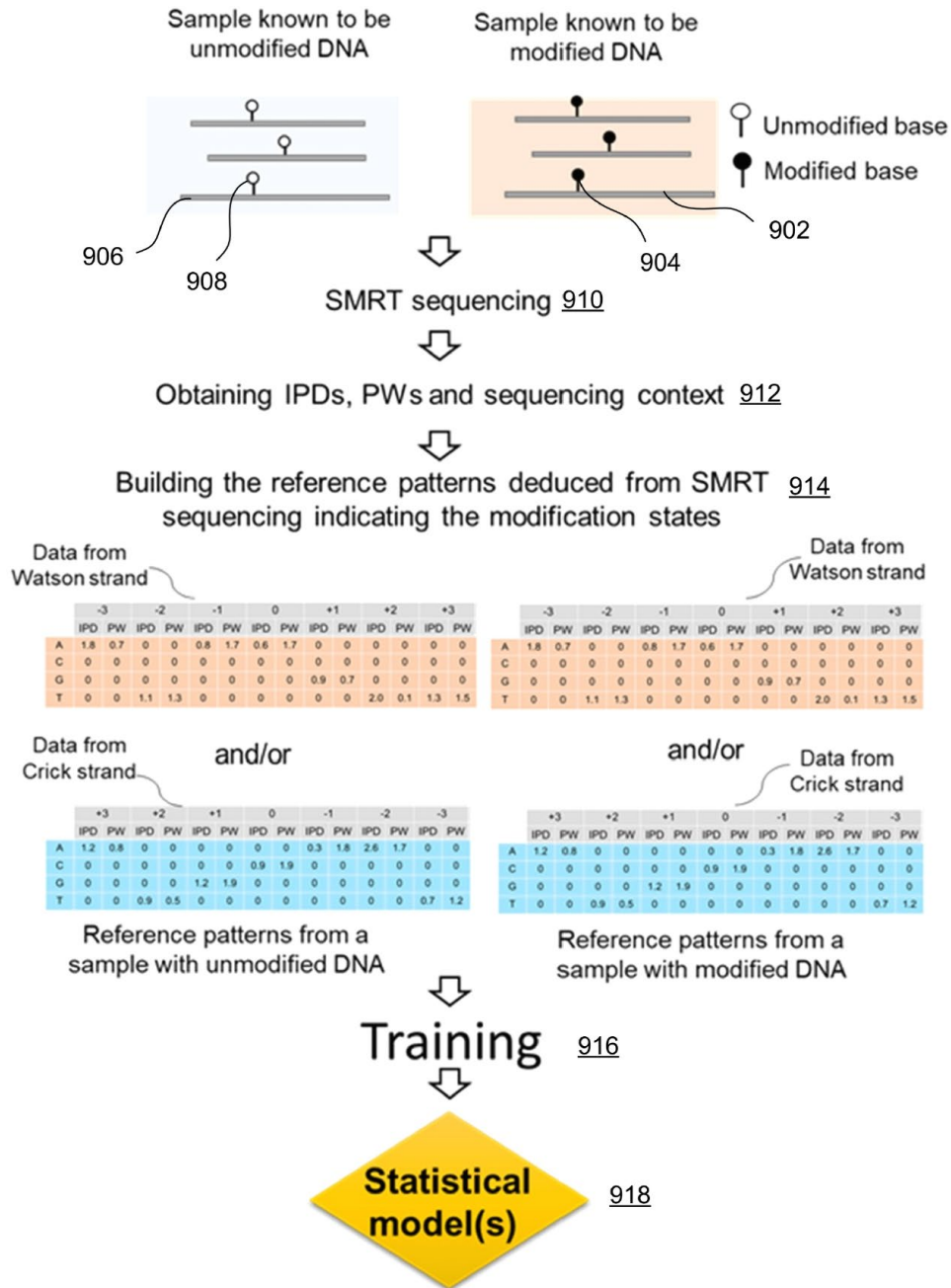
for base modification analysis. The relative positions of -1, -2, and -3 indicated the position 1-nt, 2-nt, and 3-nt, respectively, upstream of the base that was subjected to base modification analysis. The relative positions of +1, +2, and +3 indicated the position 1-nt, 2-nt, and 3-nt, respectively, downstream of the base that was subjected to base modification analysis. Each position includes 2 columns which contain the corresponding IPD and PW values. The following 4 rows (rows **408**, **412**, **416**, and **420**) corresponded to 4 types of nucleotides (A, C, G, and T) in the strand (e.g., Watson strand), respectively. The presence of IPD and PW values in the matrix depended on which corresponding nucleotide type was sequenced at a particular position. As shown in FIG. 4, at the relative position of 0, the IPD and PW values were shown in the row indicating “G” in the Watson strand, suggesting that a guanine was called in the sequence result at that position. The other grids in a column that did not correspond to a sequence base would be coded as ‘0[.]’ As an example, the sequence information corresponding to the 2-D digital matrix (FIG. 4) would be 5'-GATGACT-3' for the Watson strand. EX1001, 23:28-51.

43. The matrix illustrated in FIG. 4 of the '794 patent is an example of an input data structure using the sequence data from a Watson strand that can be used to train a model. Additional input data structures for training a model using the Crick strand and the combination of the Watson and Crick strands are illustrated in FIGs. 5-8 and use the above-described format. Supporting disclosure for each figure and matrix is found in the '794 patent, 23:52-25:41, EX1001. The matrices illustrated

in FIG. 4 and throughout the '794 patent are illustrated for the reader to understand and visualize. A person of ordinary skill the art reviewing this disclosure would understand that for a model, the matrix can be implemented in multiple ways. For example, "A, C, G and T [may be] coded with [1,0,0,0], [0,1,0,0], [0,0,1,0], and [0,0,0,1], respectively. . . ." EX1001, 35:56-57. Alternatively, the value "1" in the coding may be replace by a numeric value of a property associated with said nucleotide.

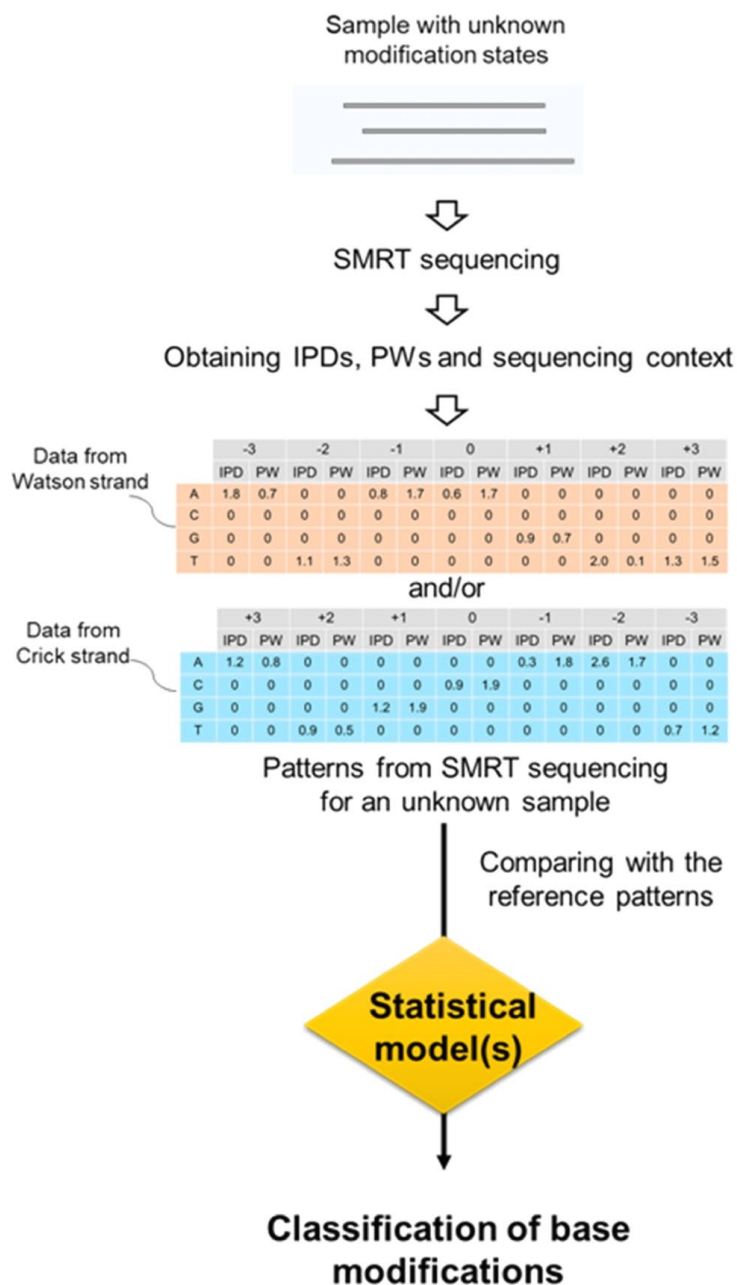
44. As further disclosed in the '794 patent, the matrices were used to train a model. FIG. 9 (below) "shows a general procedure as to how to use the measurement window to determine any base modifications." EX1001, 25:58-59. First, DNA samples known to be modified and unmodified were subjected to single molecule real-time sequencing. EX1001, 25:61-26:2. The circular DNA molecules may be sequenced multiple times generating multiple subreads and aligned to a reference genome. EX1001, 26:2-22. Then, "[a]t stage **912**, from the alignment result, IPDs, PWs, and sequence context surrounding a base that was subjected to base modification analysis were obtained." EX1001, 26:23-25. "At stage **916**, a number of 2-D matrices containing the reference kinetic patterns derived [from] molecules with known base modifications were used to train the analytical, computational, mathematical, or statistical model(s)." EX1001, 26:29-32. These

matrices are the input data structures of claim 1. “At stage **918**, a statistical model is developed resulting from the training.” EX1001, 26:32-34.



EX2001, FIG. 9.

45. The '794 patent also discloses using the same input data structure for detecting methylation (testing data) in the same format, *i.e.*, the same window with the same properties (nucleotide identity, nucleotide position relative to a target nucleotide position, nucleotide PW and IPD) as the input data structure used to train the model. For example, FIG. 10 illustrates detecting methylation using the same input data structure shown in FIG. 9 to train the model. “FIG. 10 shows a general procedure as to how the statistical model(s) learned from DNA samples that carried known states of base modifications can detect base modifications.” EX1001, 26:58-61. First, DNA samples with unknown states of base modification were subjected to single molecule real-time sequencing. EX1001, 26:61-62. The circular DNA molecules may be sequenced multiple times generating multiple subreads and aligned to a reference genome, or to each other. EX1001, 26:62-65. Then, FIG. 10 follows the similar steps as FIG. 9, “IPDs, PWs and sequence context” surrounding a base that is subject to base modification analysis are obtained to create matrices (the input data structures) containing the same reference kinetic patterns used to train the model. The model then compares the patterns from matrices of the unknown sample and to the reference patterns created by the training of the model to classify a base as modified or unmodified. FIG. 10 is shown below.



EX2001, FIG. 10.

46. FIGs. 11, 13 and 15 illustrate additional training methods using the matrices as input data structures and FIGs. 12, 14, and 16 illustrate the use of the model trained according to FIGs. 11, 13 and 15 to detect base modifications using

the matrices as input data structures. Supporting disclosure for these methods are disclosed in the '794 patent, 26:59-30:59, EX1001. Additional input structures (e.g., various 3-D, 2-D and 1-D structures) are disclosed in the '794 patent, 29:54-30:18, EX1001.

47. The '794 patent discloses the creation of a methylated library of 421,614 circular consensus sequences and an unmethylated library of 446,285 circular consensus sequences that were used to train and validate the input data structure in a convolutional neural network. EX1001, 30:64-32:17.

48. In sum, the '794 patent discloses how to train a model by using training data provided in the same format, *i.e.*, the same window with the same properties (nucleotide identity, nucleotide position relative to a target nucleotide position, nucleotide PW and IPD), as the input data structure that would be used to detect DNA modifications (testing data). The '794 patent discloses systematically probing the predictive power of the features of the input data structure of claim 1 as well as other features.

C. The '794 patent discloses probing the predictive power of features.

49. The '794 patent discloses “evaluat[ing] the power of each feature (sequence context, IPD, and PW) in predicting the methylation state of CpG by including a subset of the features in the model.” EX1001, 32:18-20. In the testing dataset (FIG. 18B), models with (i) sequence context only, (ii) IPD only, and (iii)

PW only, resulted in area-under-the-curve (AUC) values of 0.5, 0.73 and 0.85, respectively. The combination of IPD and sequence context improved performance with an AUC of 0.85. The combination of claim 1, sequence context (“Seq,” i.e., nucleotide identity and relative position to the target position), IPD, and PW substantially improved performance with an AUC of 0.94, which had a comparable performance with the training dataset (FIG. 18A). EX1001, 32:18-29. FIG. 18 illustrates the results.

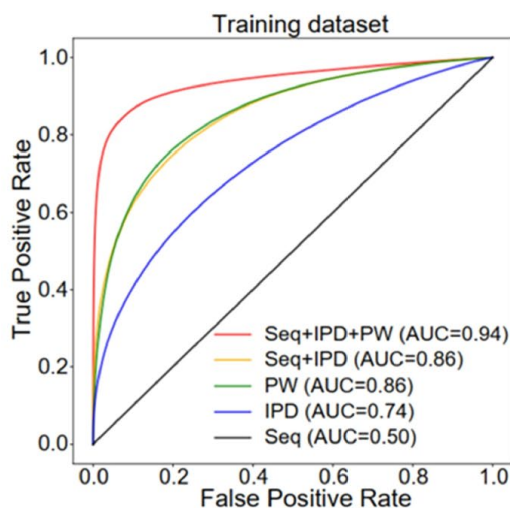


FIG. 18A

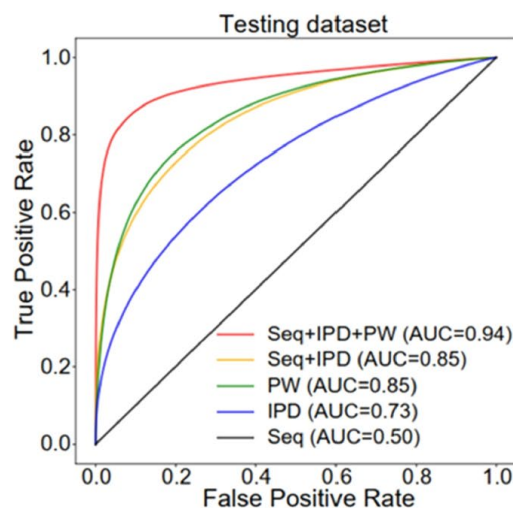
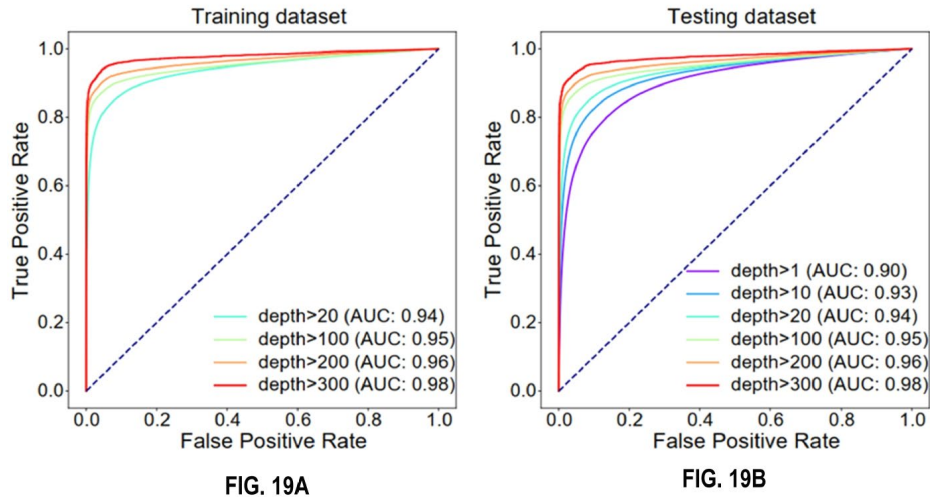


FIG. 18B

EX2001, FIG. 18.

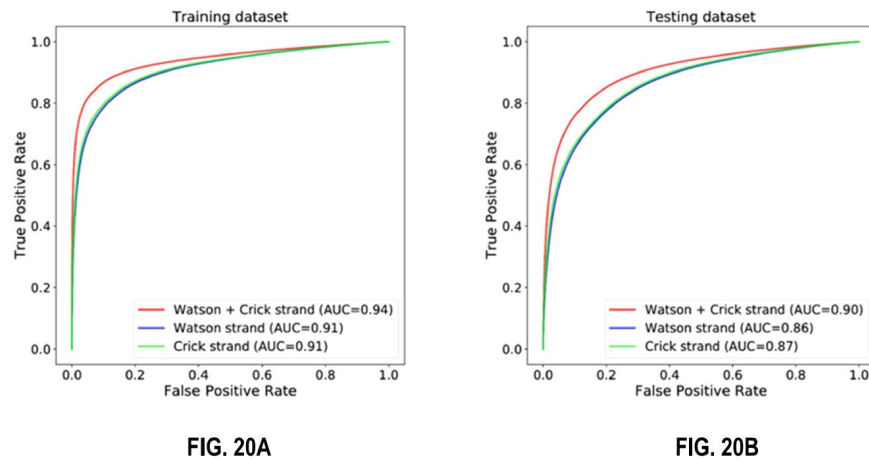
50. The '794 patent discloses evaluating the power of subread depth in predicting the methylation state of a CpG site. “We defined the subread depth of a CpG site as the average number of subreads covering it and its surrounding 10 bp. As shown in FIG. 19A and FIG. 19B, the higher the subread depth of a CpG site, the

higher the accuracy of detection of methylation we would achieve.” EX1001, 32:30-34. FIG. 19 illustrates the results.



EX2001, FIG. 19.

51. The '794 patent discloses evaluating the effect of Watson and/or Crick strands information on the performance of methylation analysis. EX1001, 32:43-56. This feature (using both strands) is claimed in claim 12 of the '794 patent. FIG. 20 illustrates the results.



EX2001, FIG. 20.

52. The '794 patent discloses evaluating the power of different numbers of nucleotides upstream and downstream of a CpG site in predicting the methylation state of a CpG site. As shown below in FIG. 21B, a window of 2 nucleotides upstream and downstream of a target nucleotide had an AUC of 0.50 which is equivalent to random guessing, whereas a 3-nucleotide window upstream and downstream of a target nucleotide had an area AUC of 0.89 or better. EX1001, 32:57-33:11.

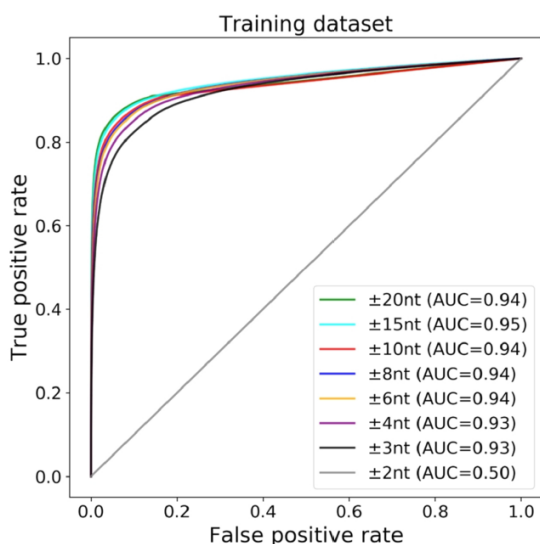


FIG. 21A

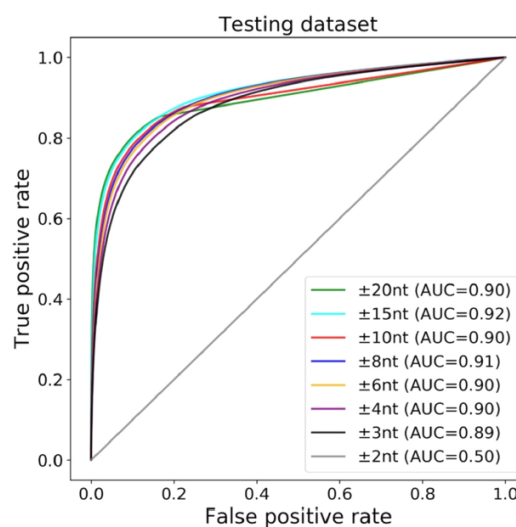


FIG. 21B

EX2001, FIG. 21.

53. The '794 patent discloses evaluating the power of different numbers of nucleotides downstream of a CpG site in predicting the methylation state of a CpG site. EX1001, 33:49-58. As shown below in FIG. 22B, a window of 2 nucleotides

upstream and downstream of a target nucleotide had an AUC of 0.50 which is equivalent to random guessing, whereas a 3-nucleotide window upstream and downstream of a target nucleotide had an area AUC of 0.87 or better.

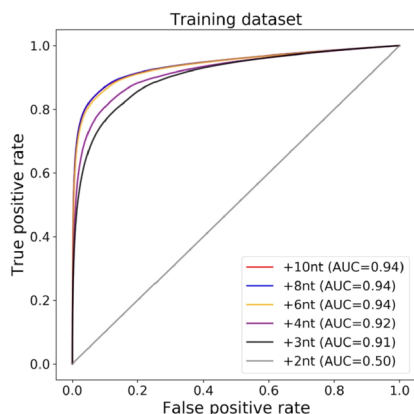


FIG. 22A

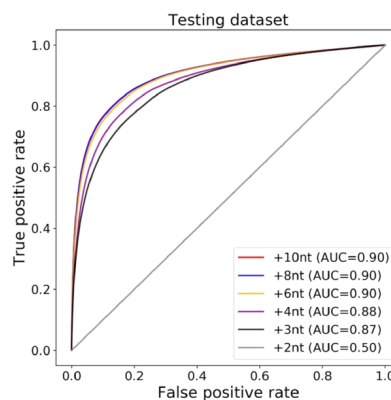


FIG. 22B

EX2001, FIG. 22.

54. The '794 patent discloses also evaluating the power of different numbers of nucleotides upstream of a CpG site in predicting the methylation state of a CpG site. EX1001, 32:59-64. As shown in FIG. 23 below using upstream nucleotide window only is equivalent to random guessing.

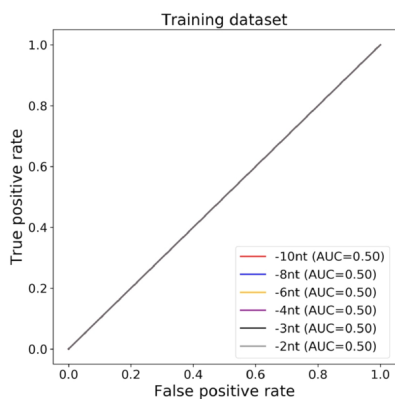


FIG. 23A

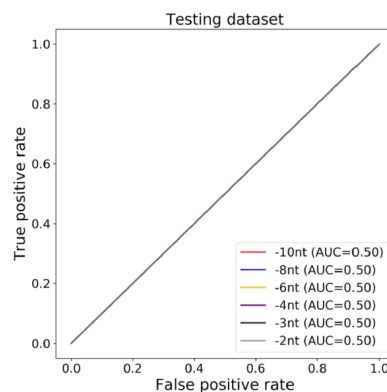
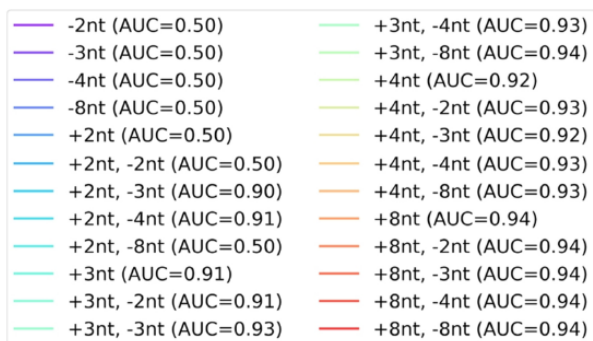
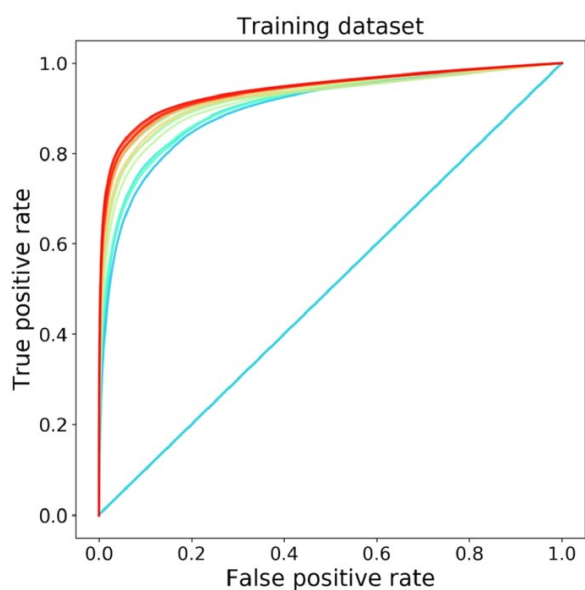


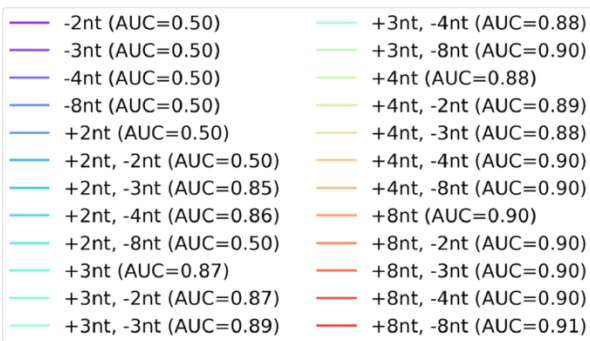
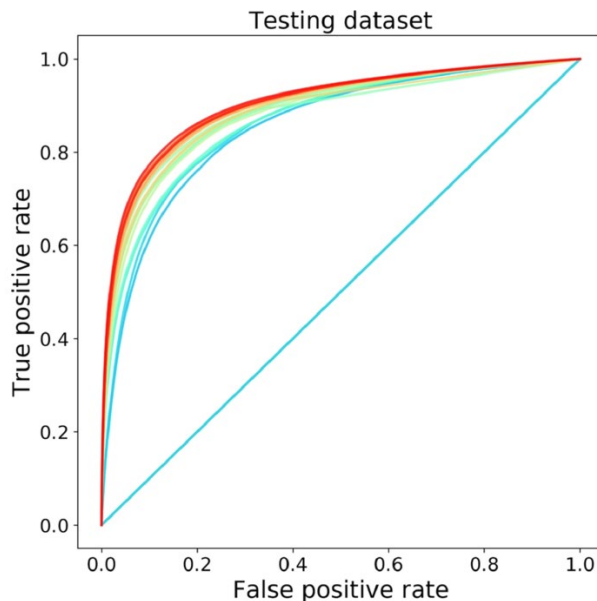
FIG. 23B

EX2001, FIG. 23.

55. The '794 patent discloses also evaluating the power of different combinations of upstream and downstream bases in predicting the methylation states. FIGs. 24 and 25 illustrate the results.



EX2001, FIG. 24.



EX2001, FIG. 25.

56. The '794 patent includes experimental data supporting practical applications of base modification detection including imprint regions in placental tissue (44:50-47:35), circulating cell free DNA from pregnant women (47:36-48:63), copy number changes in tumor tissue (49:56-51:14), maternal plasma DNA tissue

mapping (51:15-52:32), regional methylation quantification (52:33-60:17), differential methylation regions between tumor tissue and adjacent normal tissue (60:18-61:60), and methylation analysis in viral DNA (61:61-62:48). EX1001.

D. Prosecution history of claim 1

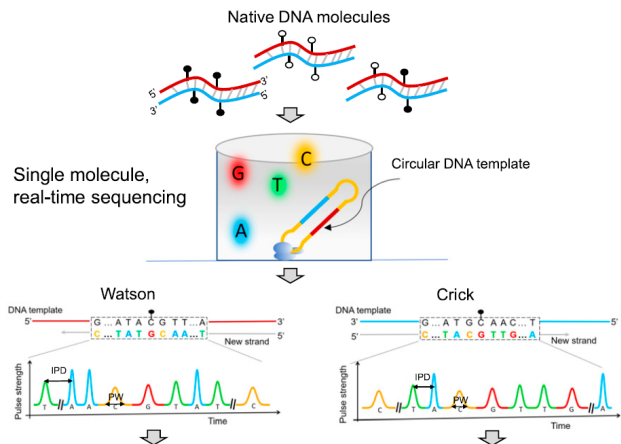
57. I have reviewed the prosecution history of the '794 patent, including the Preliminary Amendment dated August 17, 2020 (EX1004, pp. 1090-95), the Office Action mailed December 30, 2020 (EX1004, pp. 795-810), the Applicant-Initiated Interview Summary dated March 9, 2021 (EX1004, pp. 216-18), the Claim Amendments and Response to the Office Action mailed December 30, 2020 (EX1004, pp. 190-210).

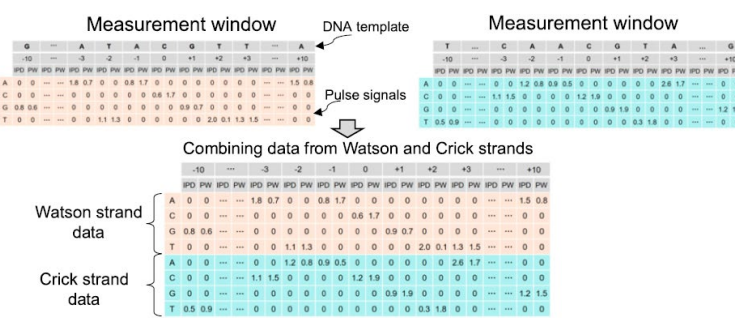
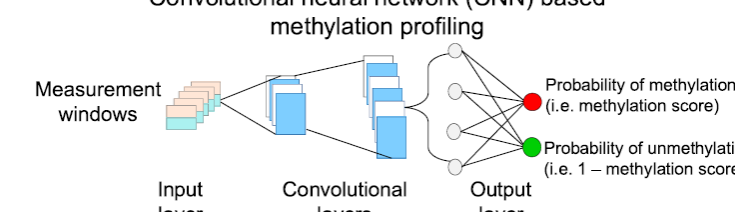
58. The Applicant-Initiated Interview Summary dated March 9, 2021 under 35 U.S.C. 102 states: "Examiner stated the method of Flusberg displays various widths of pulses in their sequence data in a window, however it does not appear that they measure the widths as a data point to be combined with interpulse duration, nucleotide identity and position in determining the presence of mutations." EX1004, pp. 217. And under 35 U.S.C. 103, it states: "See comments in 102 above." *Id.*

59. I understand that Examiner concluded Flusberg and the '338 patent do "not appear that they measure the width as a data point to be combined with interpulse duration, nucleotide identity and position in determining the presence of mutations." *Id.* I agree with Examiner's conclusion.

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

60. I have reviewed Tse et al., “*Genome-wide Detection of Cytosine Methylation by Single Molecule Real-time Sequencing*,” Proceedings of the National Academy of Sciences, 118(5) (2021) (“Tse”) (EX1030), which I understand is authored by the inventors of the ’794 patent and other collaborators. In my opinion, the Tse publication discloses an embodiment of claim 1 of the ’794 patent. Below is a claim chart mapping disclosure of the Tse publication to claim 1. The embodiment disclosed in Tse falls within the scope of claim 1 of the ’794 patent as set forth in the following table.

Claim 1 of the 794 Patent EX1030	Disclosure in Tse (EX1030)
1. A method for detecting a modification of a nucleotide in a nucleic acid molecule, the method comprising:	“Detection of Cytosine Methylation” EX1030, p. 1, Title.
(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	 The diagram illustrates the process of single molecule, real-time sequencing. At the top, 'Native DNA molecules' are shown as double helices. Below, a 'Single molecule, real-time sequencing' setup is depicted with a 'Circular DNA template' containing nucleotides G, T, and C. A 'New strand' is being synthesized. The diagram shows two examples: 'Watson' and 'Crick' strands. Each example includes a 'DNA template' sequence and a 'New strand' sequence. Below each sequence is a graph of 'Pulse strength' versus 'Time', showing peaks corresponding to the nucleotides. The Watson example shows a sequence G...ATACGTT...A and a new strand C...TATGCA...T. The Crick example shows a sequence G...ATGCAC...T and a new strand C...TACGTT...A. The graphs show peaks for each nucleotide, with the Watson graph showing a peak for 'C' and the Crick graph showing a peak for 'G'. EX1030, Fig. 1 (partial).

Claim 1 of the 794 Patent EX1030	Disclosure in Tse (EX1030)
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;	 The diagram illustrates the process of creating an input data structure for a convolutional neural network (CNN) based methylation profiling. It shows a DNA template with a target sequence (G A T A C G T T A) and a measurement window. Pulse signals are generated for each nucleotide, with values ranging from 0 to 1.5. The data is then combined from the Watson and Crick strands, resulting in a matrix of values for each nucleotide and its position within the window. EX1030, Fig. 1 (partial).
(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 first data structures corresponding to a respective window of	 The diagram shows a CNN-based methylation profiling model. It takes measurement windows as input, processes them through an input layer, convolutional layers, and an output layer. The output layer produces two probabilities: the probability of methylation (i.e. methylation score) and the probability of unmethylation (i.e. 1 - methylation score). EX1030, Fig. 1 (partial). “We used amplified DNA (unmethylated) and M.SssI-treated DNA (methylated) (M.SssI being a CpG methyltransferase) to train a convolutional

Claim 1 of the 794 Patent EX1030	Disclosure in Tse (EX1030)
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,	neural network.” EX1030, p. 1. “The measurement windows associated with methylated CpG sites (the M.SssI-treated DNA dataset) and those associated with unmethylated CpG sites (the WGA DNA dataset) were used for training the HK model through CNN. Data within each measurement window flanking a cytosine within a CpG context, including the sequence context, mean IPDs, and PWs originating from subreads across individual nucleotides, were entered into the HK model. Each target output (i.e., analogous to a dependent variable value) for a CpG site in M.SssI-treated DNA datasets was assigned as “1” while each target output for a CpG site in WGA DNA datasets was assigned as “0.” The patterns present in the measurement windows of methylated and unmethylated CpG sites were used for training CNN to determine the parameters (often called weights) of the HK model.” EX1030, p. 10. “The HK model was applicable to data produced by different sequencing kits and sequencers as long as the training and testing processes were based on the same experimental conditions.” EX1030, p. 4.

Claim 1 of the 794 Patent EX1030	Disclosure in Tse (EX1030)
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.	Convolutional neural network (CNN) based methylation profiling Measurement windows Input layer Convolutional layers Output layer Probability of methylation (i.e. methylation score) Probability of unmethylation (i.e. 1 - methylation score) EX1030, Fig. 1 (partial). Also see EX.1030, Fig. 2D showing accuracy characterized by AUC to be above 9.3. D Testing dataset True positive rate False positive rate Sequel I kit 3.0 (AUC: 0.97) Sequel II kit 1.0 (AUC: 0.96) Sequel II kit 2.0 (AUC: 0.93) EX1030, Fig. 2 (partial).

VIII. Flusberg does not teach or suggest to a person of ordinary skill in the art the “input data structure” of claim 1.

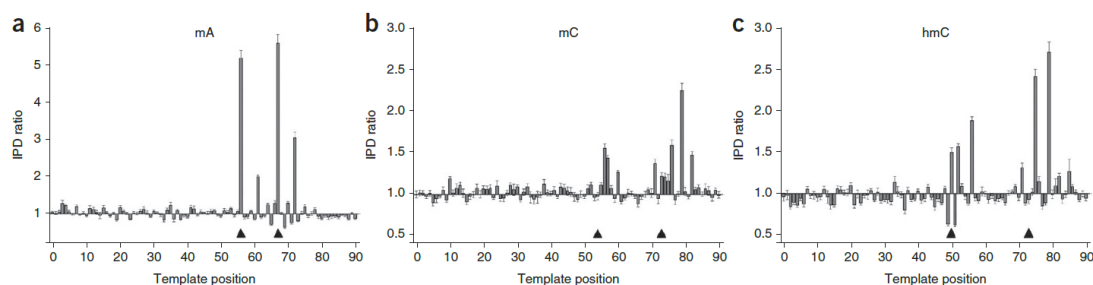
61. Flusberg is a proof-of-concept paper that shows that the presence of methylated DNA affects polymerase kinetic data recorded during SMRT sequencing. Flusberg hypothesized “that methylated bases in a DNA template might

be detected directly on the principle that their presence affects polymerase kinetics during SMRT sequencing (**Fig. 1**).” EX1006, p. 462. To test this hypothesis Flusberg “designed several synthetic DNA templates that were identical except for their methylation status at specific sites.” *Id.* In the ’338 patent, Flusberg et al. characterized these synthetic DNA templates as “contrived patterns of methylated and unmethylated bases, along with control sequences having the same primary sequence but without any methylation.” EX1007, 75:35-38.

62. Flusberg’s “control template contained no methylation, and other templates contained several mA, mC or hmC bases.” EX1006, p. 462. The sequences of those four synthetic DNA templates are set forth on page 479 of EX1006: Supplementary Note: Sequences of synthetic DNA templates. The synthetic sequences were constructed based upon ~35-base single stranded DNA oligonucleotides for which the base modifications were verified by mass spectrometry. *Id.*, p. 466 (Preparation of DNA templates).

63. As disclosed in Flusberg and shown on page 479 “[i]n all cases, mA was located in a GATC sequence context, and mC and hmC were located in C-G contexts.” EX1006, p. 462. In this passage, the term “sequence context” does not refer to input data, but rather the known window around a base (adenine) modification. Flusberg’s kinetic data analysis focused on IPD ratios that correlated with the location of known methylated bases. “We sequenced these templates and

then compared the average IPD in each of the methylated templates to the average IPD in the control template by computing their ratio at every template position.” *Id.*, Fig. 2 of Flusberg shows the ratios of the average IPD of the methylated template to the average IPD of the control template, plotted by DNA template position. The control and methylated templates are identical except for two methylated sites shown as triangles in Fig. 2.



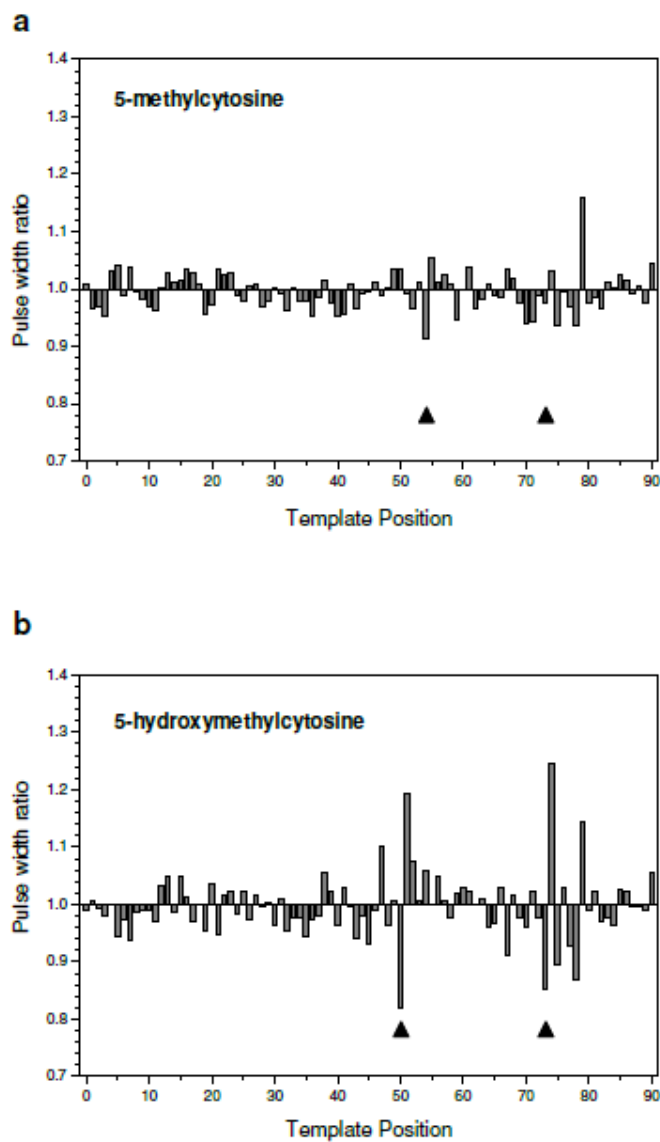
EX1006, Fig. 2.

64. Flusberg noted that the kinetic impact of methylation was not restricted to the nucleotide being incorporated opposite the modified base. The modified base had an upstream and downstream kinetic effect on IPD. Because the IPD ratio patterns differed between the two methylated bases and “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**),” Flusberg concluded that “in general, the kinetic signatures of methylation will be sequence context-dependent.” EX1006, pp. 462-63. Again, the use of the term sequence

context in Flusberg is not as input data, but rather an observation regarding pattern of IPD ratios associated with the nucleotide positions surrounding a known methylation site, *i.e.*, a window.

65. Flusberg also measured the pulse width ratios and found that “[p]lots of pulse width ratios had more pronounced excursions for hmC than for mC (**Supplementary Fig. 1**).” EX1006, p. 463.

Supplementary Figure 1: Pulse width ratios for mC and hmC



EX1006, Supplementary Fig. 1.

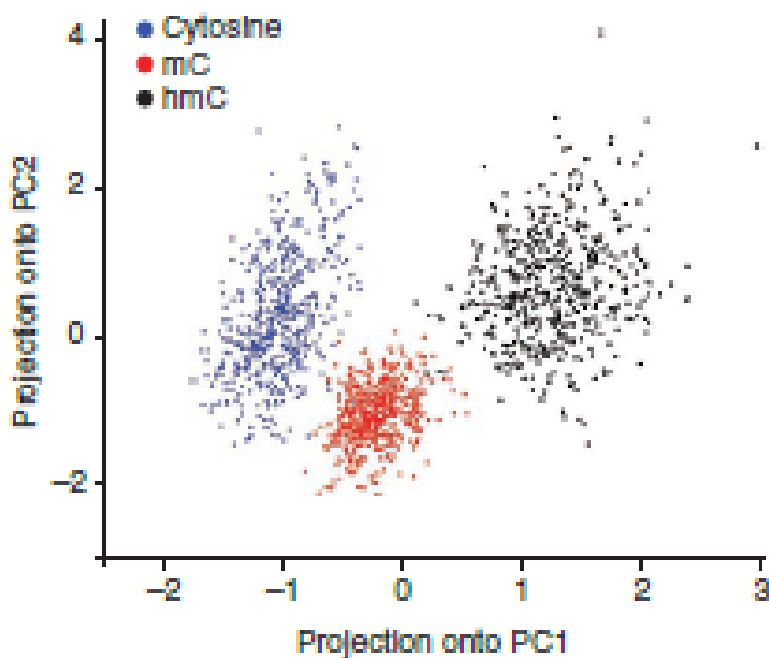
66. Notably, the pulse width ratios were significantly lower than the IPD ratios, and the IPD ratio also exhibits substantially higher signal-to-noise ratio than

the pulse width. The IPD ratios had a dynamic range of about 2.3 around a methylated cytosine and about 2.8 around a hydroxymethylated cytosine, while the background levels had a dynamic range of about 0.2 (Figs. 2b & c). The pulse width ratios had a dynamic range of about 0.4 around the methylated cytosine and about 0.5 around the hydroxymethylated cytosine while the background levels had a dynamic range of about 0.1 (Supplementary Figs. 1a & b). In other words, IPD ratios had a substantially higher signal-to-noise ratio (2-3X) defined by the ratio of the dynamic range around a modified base compared to background positions.

67. A person of ordinary skill in the art reviewing these data would conclude that IPD exhibits substantially higher signal-to-noise ratio and thus much greater sensitivity to a modified base than pulse width. Subsequent studies of PacBio and its collaborators focused on IPD and did not consider PW in their analysis, as discussed in paragraph 138. *See* Schadt (EX1010), where Flusberg's co-authors Jonas Korlach and Tyson A. Clark explained: "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." EX1010, p. 130.

68. Flusberg reasoned that "because each modification had a unique IPD and pulse width signature in a given context (position 73, for example, **Figs 2b & c**

and **Supplementary Fig. 1**), this approach opens up the intriguing possibility of directly distinguishing between cytosine, mC and hmC during real-time sequencing.” EX1006, p. 463. Flusberg “used principal component analysis to find the combination of weights for the IPD and pulse width signals at the various template positions near the putative modification that optimizes the resolution (**Supplementary Table 1**).” *Id.* Flusberg then plotted those kinetic signatures in Fig. 3.



EX1006, Fig. 3.

69. Fig. 3 illustrates that using principal component analysis, IPD and PW ratios of the bases 2-nt upstream and 6-nt downstream (71-79) around position 73 of the template DNA, the three cytosine species, unmethylated cytosine, mC and hmC can be separated from each other. While Fig. 3 appears to separate the three cytosine species using principal component analysis, the experiment is idealized using contrived DNA templates having an identical primary DNA sequence with an unmethylated control template and a methylated template having two cytosine methylation sites, of which one was analyzed using principal component analysis.

70. When the inventors of the '794 patent tested the principal component method of Flusberg using 5,948,348 DNA segments which were 2-nt upstream and 6-nt downstream of a cytosine that was subjected to methylation analysis, the signals deduced from the principal component analysis with the use of IPDs and PWs were found to be largely overlapping between methylated cytosines (mC) and unmethylated cytosines (C) fragments showing that the method described by Flusberg could not differentiate 5-methylcytosine from unmethylated cytosine. EX2001, FIG. 28. The '794 patent also shows that the principal component method of Flusberg resulted in an AUC of 0.55 which is essentially random guessing. EX2001, FIG. 29.

71. The next section of the Flusberg Results: Methylation detection by circular consensus sequencing (EX1006, pp. 463-64) shows that repeated measurements of IPD at a particular template position allowed for better separation of methylated and unmethylated adenine. Fig. 4 illustrates this separation.

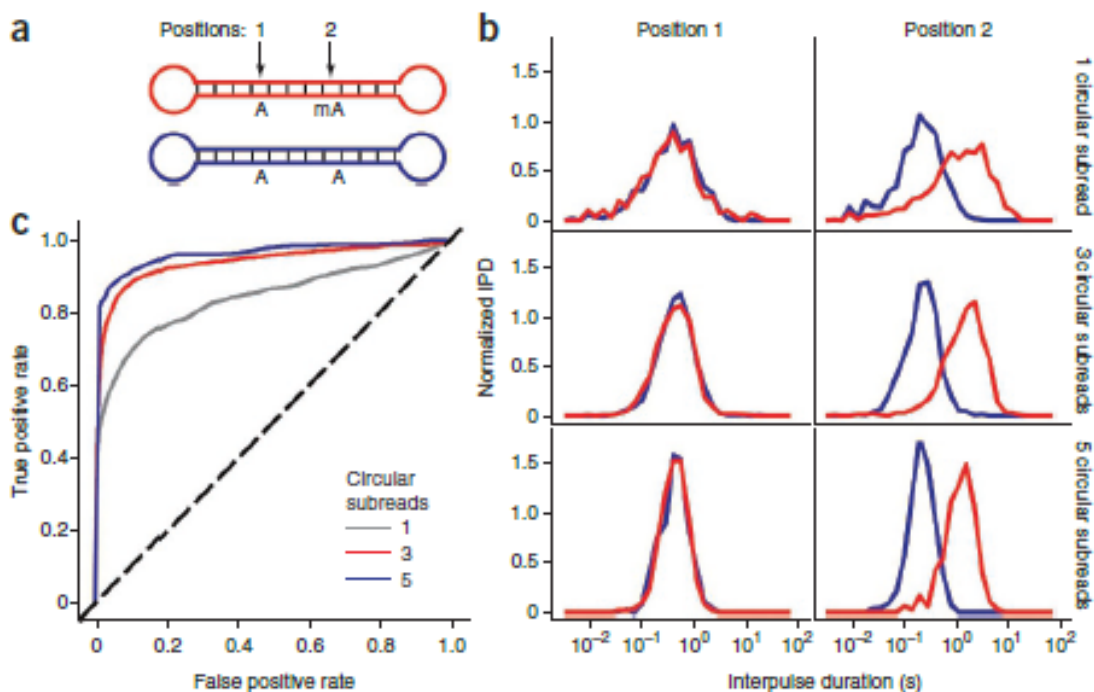


Figure 4 | IPD distributions for adenine and mA in synthetic DNA templates. (a) Schematic of the 199-base DNA templates used in this experiment. (b) IPD distributions at the indicated positions for both templates, colored as in a. The histograms depict the distributions of mean IPD (averaged over the indicated number of circular subreads). (c) Receiver operating characteristic curves, based on the IPD distributions from the differentially methylated position in b and parameterized by IPD threshold, for assigning a methylation status to an adenine nucleotide after 1, 3 and 5 circular consensus sequencing subreads. The dashed line depicts the receiver operating characteristic curve for randomly guessing the methylation status.

EX1006, Fig. 4.

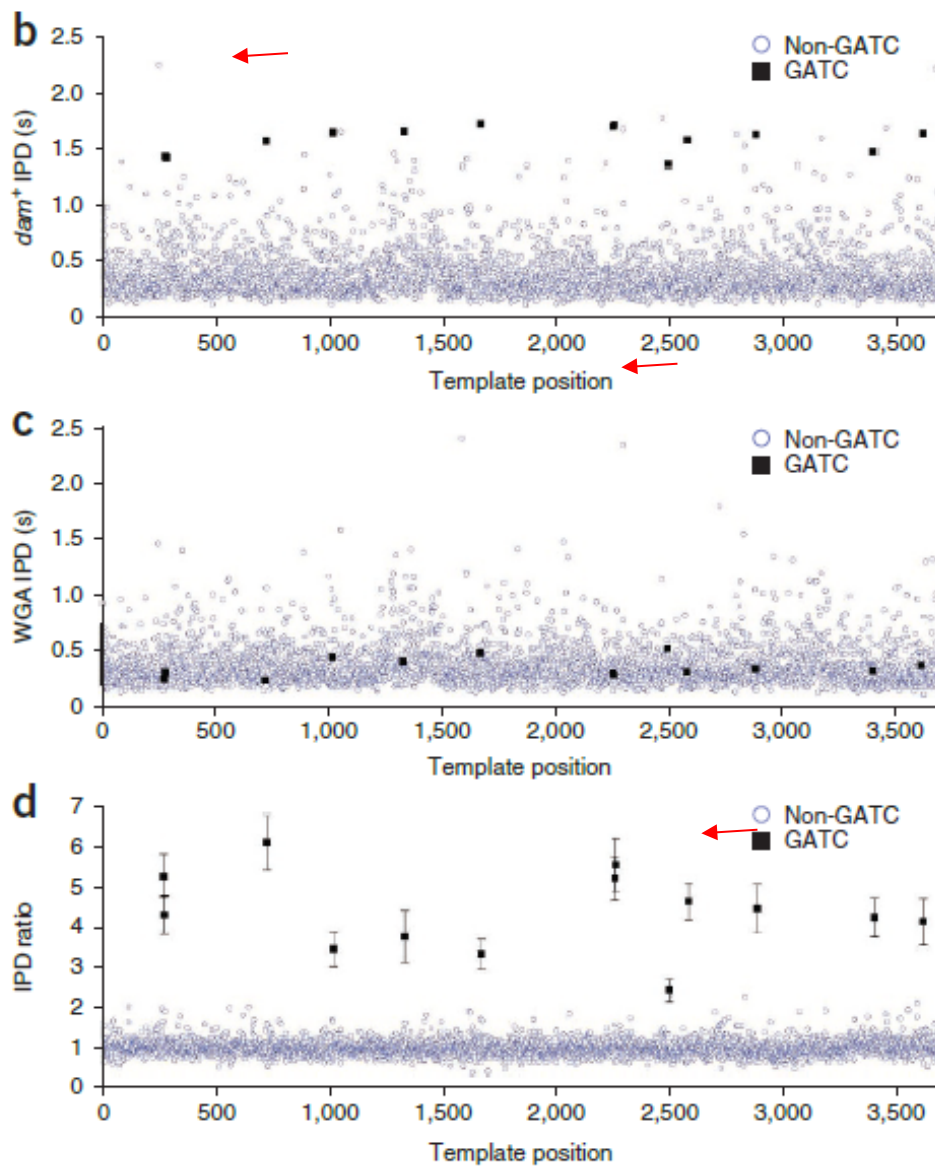
72. In contrast to FIG. 29 of the '794 patent, Flusberg Fig. 4c ROC curves address the number of circular consensus sequencing reads and shows that more

reads are better at separating methylated from unmethylated adenine at a particular position using IPD.

73. The next section of Flusberg, Results: Adenine methylation in *Escherichia coli*, addressed methylation of genomic DNA. EX1006, pp. 464-65. In this experiment, Flusberg used a *Caenorhabditis elegans* fosmid from a DNA adenine methyltransferase positive (*dam*⁺) *E. coli* strain for the methylated sequences and used whole-genome amplification (WGA) of a portion of the fosmid sample to create the unmethylated DNA control templates. Then, using SMRT sequencing, Flusberg compared the IPD kinetics between the unmethylated and methylated sequences. EX1006, p. 464.

74. Flusberg observed that average IPD values of methylated motifs (GATC, black squares) in methylated DNA samples (*dam*⁺, Fig. 5b) were higher (~1.5 s) than the average IPD for the same motifs in the unmethylated control sample (WGA, Fig. 5c) (<~0.5 s). A few unmethylated motifs (Non-GATC, open circles) showed IPD values comparable or even higher than methylated motifs (e.g., >~2.0 s, red arrows) in both the methylated DNA samples (*dam*⁺, Fig. 5b) and the unmethylated DNA samples (WGA, Fig. 5c). To eliminate the noise caused by the relatively high IPD values of unmethylated motifs, Flusberg used IPD ratios obtained by comparing IPD values of methylated DNA samples over IPD values of

the corresponding unmethylated DNA samples with identical sequences. Flusberg, Fig. 5(d), which is obtained by dividing Fig. 5(b) (IPD values of the *dam*⁺ sample) by Fig. 5(c) (IPD values of the WGA samples) of the corresponding positions, respectively. EX1006, p. 464.



EX1006, Fig. 5.

75. Flusberg teaches using IPD ratios over a control template to reduce noise caused by non-methylated IPD. As Flusberg's co-authors Jonas Korlach and Stephen W. Turner explain in Suzuki (EX2005), IPD ratios "eliminate the context-dependent fluctuation of the IPD values." EX2005, p. 2912. A person of ordinary skill in the art would be discouraged from following the path of using nucleotide identity as input data to detect base modification because Flusberg teaches using IPD ratios as a normalization to remove the contribution of nucleotide identity. A person of ordinary skill in the art would be led in the direction of IPD ratios which is divergent from the path followed by and claimed in the patent - the combination of nucleotide identity, IPD and PW.

76. Table 1 shows that the IPD ratios at various GATC motifs ranged from 2.43 to 6.12.

Table 1 | IPD ratios for each fosmid GATC motif

Position	Sequence	IPD ratio	P value
273	TGCCATGATCTAGATC	5.28	2.84×10^{-18}
279	GATCTAGATCATCGTG	4.32	4.41×10^{-16}
720	TTCTATGATCAGGGAG	6.12	7.00×10^{-21}
1015	GCGTGGGATCTGTATG	3.46	6.25×10^{-13}
1329	TATCACGATCTCATT	3.78	2.35×10^{-12}
1668	TAGTTGGATCAAGAGA	3.33	2.61×10^{-13}
2256	CTTTTGGATCAGATCC	5.22	1.70×10^{-10}
2261	GGATCAGATCCAATTA	5.56	1.43×10^{-22}
2499	CAGATGGATCAATCAA	2.43	4.44×10^{-7}
2583	ATTTTGGATCTAGTTT	4.65	2.00×10^{-21}
2887	ATTCGCGATCTCCACA	4.46	1.57×10^{-14}
3402	CCTCAAGATCATCATC	4.24	2.04×10^{-15}
3619	GCCAGCGATCATATTT	4.13	2.05×10^{-10}

EX1006, Table 1.

77. Notably, Flusberg did not call any of the adenine positions in the GATC motif as methylated. Flusberg just measured the average IPD values for an expected methylated position and reported a ratio. This is another proof of principle study where Flusberg showed the sequence context could have an influence on the IPD ratios around the GATC, but did not teach nucleotide identity as an input data. And with respect to the local sequence context in **Supplementary Fig. 5** Flusberg confirmed that “[t]he ‘local context’ of detected bases was defined as the two bases previously detected, the detected base itself and the next base detected afterward.” EX1006, p. 467. Here, Flusberg uses “local context” to refer to a four-base motif that was under investigation, not using nucleotide identity as input data.

78. The Flusberg experiments used DNA control templates that are identical but for the base(s) methylated to enable the calculation of IPD ratios, which serve to mitigate the effects of sequence identities on IPD values. The IPD ratio is essentially a normalization that removes the contribution of nucleotide identity. Consequently, the identity of the nucleotides of the template sequences is no longer a variable, but rather a constant. Dr. Dessimoz’s declaration explains that Flusberg does not disclose nucleotide identity as an input feature and I agree. “Flusberg’s specific example avoids using sequence context as an express input variable because it is held constant – Flusberg’s principal component analysis model is trained on a particular sequence that is varied in methylation state, but not in nucleotide

sequence.” EX1002, p. 63, ¶107. There is no teaching or suggestion to a person of ordinary skill in the art in Flusberg to use nucleotide identity as input data.

IX. The '338 patent does not teach or suggest to a person of ordinary skill in the art the “input data structure” of claim 1.

79. U.S. Patent No. 9,175,338 titled “Methods for Identifying Nucleic Acid Modifications” issued November 3, 2015 and was filed as U.S. Patent Application No. 12/945,767 on November 12, 2010 (“the '338 patent”). EX1007. While much of the disclosure of the '338 patent is prophetic, the experimental data disclosed in the Examples section and the referenced Figures is duplicative of Flusberg’s observations that base modifications affect polymerase kinetics.

80. The disclosure of the '338 patent covers a variety of topics including the single molecule sequencing technique (Section II, EX1007, 17:35-24:45), polymerase modification for single molecule sequencing (Section III, EX1007, 24:46-26:60), secondary structure detection (Section IV, EX1007, 26:61-27:63), modified and non-natural nucleotide analogs and base pairing (Section V, EX1007, 27:64-30:62), chemical modification of template (Section VI, EX1007, 30:63-42:7), detection of agent-nucleic acid interactions (Section VII, EX1007, 42:8-51:39), nucleic acid binding agents as analytical tools (Section VIII, EX1007, 51:40-61:5), data analysis (Section IX, EX1007, 61:6-69:51), systems (Section X, EX1007, 69:52-74:19), bar coding applications (Section XI, EX1007, 74:20-55),

metagenomic identifications (Section XII, EX1007, 74:56-75:30) and examples (Section XIII, EX1007, 75:31-80:56).

81. Section XIII, Examples of the '338 patent discloses the experiments and the results. The Examples section is broken down into subsections. The first subsection is titled "Detection of 5-methylcytosine (5-MeC)." EX1007, 75:32-76:64. Similar to Flusberg, the '338 patent discloses that for detection of 5-methylcytosine (5-MeC), SMRT sequencing "was performed on short, synthetic DNA oligos with contrived patterns of methylated and unmethylated bases, along with control sequences having the same primary sequence but without any methylation."¹ EX1007, 75:34-38. Because the 5-methylcytosine experiments used an identical DNA sequence, the nucleotide identity of the sequence is not a variable, but rather a constant.

82. According to the '338 patent, "[t]hese templates provided unequivocal fluorescence pulse patterns and tempos that demonstrated how the combination of sequence context and methylation status affected interpulse duration." EX1007, 75:38-41. Like Flusberg the '338 patent's reference to sequence context is

¹ In contrast to Flusberg, the 5-methylcytosine (5-MeC) example does not disclose what the contrived DNA template sequences are or how they were generated. EX1007, 75:32-76:64.

observational - that the IPD around a modified base will differ several bases upstream and downstream of the modified base, *i.e.*, a window. A person of ordinary skill in the art would understand that sequence context in this example of the '338 patent does not refer to nucleotide identity because in addition to the nucleotide sequences of the template being undisclosed, the nucleotide identity of the template sequences is a constant in this experiment, not a variable.

83. FIGs. 11-13 illustrate that methylated DNA affects IPD recorded during SMRT sequencing at and around the methylated base. “The difference in average interpulse durations between the two templates was visible both at the 5-MeC position and in the vicinity of the 5-MeC position.” EX1007, 75:44-46. “Further, methylcytosine was shown to have an effect on interpulse duration (IPD) not only at the methylated base, but over a range of several bases upstream and downstream of the position of the methylcytosine. EX1007, 76:40-43. “That data clearly showed that IPD was increased in the region comprising the methylated bases.” EX1007, 76:56-58.

84. While the detection of 5-methylcytosine (5-MeC) example includes new FIGs. 11, 12 and 13, they are duplicative of the Flusberg experiments. The experimental support and the data shown in FIGs. 11a and 11b are duplicative of Figs. 4a and 4b of Flusberg. EX1006, p. 463. The supporting experiments and that data shown in Figs. 4a, 4b and 4c of Flusberg demonstrate that additional subreads

from SMRT sequencing runs allowed better separation of distributions of mean IPD for methylated and unmethylated bases. *Id.* The difference between FIGs. 11a and 11b of the '338 patent and FIGs. 4a and 4b of Flusberg is the methylated nucleotide. FIGs. 4a and 4b of Flusberg analyzed a methylated adenine (6-mA), whereas FIGs. 11a and 11b of the '338 patent analyzed a methylated cytosine (5-mC). Notably, the '338 patent does not include receiver operating characteristics curves for methylated cytosine as shown in EX1006, Fig. 4c. Consequently, the '338 patent did not provide a quantitative evaluation of how meaningful the separation was.

85. FIG. 12 shows “the ratio of the average IPD in the methylated template to the average IPD in the unmethylated template, plotted versus DNA template position” for 5-methylcytosine. EX1007, 76:46-49. Similarly, Fig. 2b of Flusberg shows “the ‘[r]atios of the average IPD in the methylated template to the average IPD in the control template, plotted versus DNA template position” for 5-methylcytosine. EX1006, p. 462. While the data does not appear identical, it is duplicative of what was shown in Flusberg, namely that the 5-methylcytosine modifications have an upstream and downstream effect on IPD.

86. FIG. 13 shows “the ratio of IPD for a different methylated template vs. an identical but unmethylated template as a function of position. Seven cytosines (shown in crosshatching) were differentially methylated (5-MeC) between the two templates.” EX1007, 76:52-56. Similarly, Flusberg Supplementary Fig. 6 shows

“the ratio of the average IPD in the methylated template to the average IPD in the control template, plotted versus DNA template position. In the region shown, the two templates are identical except at the seven positions marked by triangles. . . .” EX1006, p. 477. FIG. 13 in the ’338 patent appears very similar if not identical to Supplementary Fig. 6 in Flusberg. FIG. 13 like FIG. 16 shows that 5-methylcytosine has an upstream and downstream effect on IPD.

87. In sum, the 5-methylcytosine experiments of the ’338 patent are largely duplicative of Flusberg. And, like Flusberg, the 5-methylcytosine experiments of the ’338 patent are simple idealized experiments using DNA sequences with contrived patterns of methylated and unmethylated bases along with control sequences having the same primary sequence but without any methylation. IPD was measured at a known methylation site and surrounding bases of a DNA template and then compared to an unmethylated control DNA template. Those experiments showed a correlation between IPD ratios and a 5-methylcytosine modification at the site of the modification and at bases upstream and downstream of the modification. However, the experiments did not show any *de novo* detection of methylation, i.e., an unknown methylated base.

88. The next Examples subsection is titled “Detection of N⁶-methyladenosine (N⁶-MeA).” EX1007, 76:65-77:62. These experiments used “similarly constructed template nucleic acids” as the contrived 5-methylcytosine

templates. EX1007, 76:66-77:1.² The N⁶-methyladenosine experiments are duplicative of Flusberg, as shown by FIGs. 14A, 14B, and 14C of the '338 patent which appear identical to Figs. 4a, 4b and 4c of Flusberg, and FIG. 15 of the '338 patent which appears identical to Fig. 2a of Flusberg. Similar to the results for 5-methylcytosine, FIG. 15 illustrates that methylated DNA affected IPD recorded during SMRT sequencing at and around the methylated base. "Like methylcytosine, methyladenosine was also shown to have an effect on IPD over a range of several bases upstream and downstream of the position of the methyladenosine." EX1007, 77:52-54.

89. Like Flusberg, the methyladenosine experiments of the '338 patent are simple idealized experiments using DNA sequences with contrived patterns of methylated and unmethylated bases along with control sequences having the same primary sequence but without any methylation. Because the N⁶-methyladenosine experiments used identical DNA templates, the nucleotide identity of the sequence is a constant, not a variable. IPD was measured at a known methylation site and

² Similar to the 5-methylcytosine templates, this example does not disclose what the contrived DNA template sequences are or how they were generated, except that they are identical but for two methylated bases in the GATC context. EX1007, 77:3-7.

surrounding bases of a DNA template and then compared to an unmethylated control DNA template. Those experiments showed a correlation between IPD ratios and a N⁶-methyladenosine modification at the site of the modification and at bases upstream and downstream of the modification. However, the experiments did not show any *de novo* detection of methylation, i.e., an unknown methylated base.

90. The next Examples subsection is titled “Detection of 5-hydroxymethylcytosine.” EX1007, 77:63-78:17. “Similar to 5-MeC and N⁶-MeA, 5-hydroxymethylcytosine was also tested and shown to have an effect on IPD over a range of several bases upstream and downstream of the position of the 5-hydroxymethylcytosine.” EX1007, 77:64-67. “The two templates are identical except for the hydroxymethylcytosine in the hydroxymethylated template...” EX1007, 78:7-8. Because the 5-hydroxymethylcytosine experiments used identical DNA templates, the nucleotide identity of the sequence is a constant, not a variable.

91. Similar to 5-methylcytosine and N⁶-methyladenosine, FIG. 16 illustrates that 5-hydroxymethylcytosine affected IPD recorded during SMRT sequencing at and around the methylated base. The IPD ratios at and around the modified base had a dynamic range of about 2.3 (about 0.5 to about 2.8) while background levels had a dynamic range of 0.3. FIG. 17 illustrates that 5-hydroxymethylcytosine affected PW recorded during SMRT sequencing at and around the methylated base. The PW ratios at and around the modified base had a

dynamic range of about 0.4 (about 0.8 to about 1.2) with background levels had a dynamic range of 0.12. In other words, the IPD ratio had a substantially higher signal-to-noise ratio (3X) defined by the ratio of the dynamic range around a modified base compared to background positions. A person of ordinary skill in the art reviewing these data would conclude that IPD exhibits substantially higher signal-to-noise ratio and thus much greater sensitivity to a modified base than PW. *See* Schadt (EX1010), where Flusberg's co-authors Jonas Korlach and Tyson A. Clark explained: "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." EX1010, p. 130.

92. The '338 patent 5-hydroxymethylcytosine experiments are duplicative of Flusberg as shown by FIG. 16 of the '338 patent which appears identical to Fig. 2c for Flusberg and FIG. 17 of the '338 patent which appears identical to Supplementary Fig. 1b of Flusberg. EX1006, 462, 469. Like Flusberg, the 5-hydroxymethylcytosine experiments of the '338 patent are simple idealized experiments using DNA sequences with contrived patterns of methylated and unmethylated bases along with control sequences having the same primary sequence but without any methylation. Those experiments showed a correlation between IPD ratios and a 5-hydroxymethylcytosine modification at the site of the modification

and at bases upstream and downstream of the modification. However, the experiments did not show any *de novo* detection of methylation, i.e., an unknown methylated base.

93. The next Examples subsection is titled “Detection of 8-oxoguanosine (8-oxoG).” EX1007, 78:18-43. This experiment, like the previous experiments, used contrived DNA templates that “are identical except for the 8-oxoguanosine in the modified template. . . .”³ EX1007, 78:28-29. Because the only difference between the templates is the 8-oxoguanosine modification, the nucleotide identity of the template sequence is a constant, not a variable.

94. FIG. 18 shows “the ratio of the average IPD in the 8-oxoguanosine template to the average IPD in the unmodified template, plotted versus DNA template position.” EX1007, 78:25-27. “These data showed that, compared to G, 8-oxoG altered IPD significantly over a window of ~10 neighboring bases surrounding the 8-oxoG position. Some positions saw an increase in IPD by a factor of as much as 6.5X.” EX1007, 78:30-33. FIG. 19 shows the “pulse width ratio (pulse width for an 8-oxoG template divided by pulse width for template with no 8-oxoG) vs. template position where the modified positions comprise 8-oxoG bases.”

³ Similar to the 5-methylcytosine templates, this example does not disclose what the contrived DNA template sequences are or how they were generated.

EX1007, 78:36-39. Like the 5-hydroxymethylcytosine PW data (FIG. 17) the PW ratios had a dynamic range at most about 0.8 (about 0.5 to about 1.3) around the modified base while the background levels had a dynamic range of about 0.12. On the other hand, the IPD ratios around the base modification had a dynamic range of about 6.2 (about 0.4 to about 6.6), while the background levels remained much lower (< about 0.1). This indicates that the IPD ratios exhibited a substantially higher signal-to-noise ratio defined by the ratio of the dynamic range around a modified base compared to background positions. A person of ordinary skill in the art reviewing this data would conclude that IPD exhibits substantially higher signal-to-noise ratio and thus much greater sensitivity to a modified base than PW. And, the '338 patent joint-inventors Jonas Korlach and Tyson A. Clark confirmed the same in Schadt "we restrict attention to IPDs." EX1010, p. 130.

95. The 8-oxoguanosine experiments of the '338 patent are simple idealized experiments using DNA sequences with contrived patterns of methylated bases along with control sequences having the same primary sequence but without any methylation. IPD was measured at a known methylation site and surrounding bases of a DNA template and then compared to an unmethylated DNA template. Those experiments showed a correlation between IPD ratios and an 8-oxoguanosine modification at the site of the modification and at bases upstream and downstream

of the modification. However, the experiments did not show any *de novo* detection of methylation, i.e., an unknown methylated base.

96. The next Examples subsection is titled “Detection of Ribonucleosides within DNA.” EX1007, 78:44-79:19. In this experiment, DNA templates were modified with two identical RNA ribonucleosides (adenosine, cytidine, uridine, or guanosine). FIG. 20 shows the sequences of the modified and unmodified DNA templates. Because the RNA ribonucleosides experiments used identical DNA templates, the nucleotide identity of the template sequence is a constant, not a variable. FIGs. 21-24 show the ratio of the average IPD in various ribonucleoside templates to the average IPD of control templates plotted vs. DNA template position. EX1007, 79:8-12. Similar to the 8-oxoG experiments, the “data showed that, compared to deoxyribonucleoside, a ribonucleoside altered IPD significantly over a window of ~10 neighboring bases surrounding the modified position. Some positions saw an increase in IPD by a factor of as much as 6.5X.” EX1007, 79:14-18.

97. The ribonucleoside experiments of the '338 patent are simple idealized experiments using DNA sequences with contrived patterns of bases and ribonucleosides along with control sequences having the same primary sequence but without any ribonucleoside. IPD was measured at a known modification site and surrounding bases of a DNA template and then compared to an unmodified DNA

template. Those experiments showed a correlation between IPD ratios and a ribonucleoside modification at the site of the modification and at bases upstream and downstream of the modification. However, the experiments did not show any *de novo* detection of ribonucleoside modification in a DNA sequence, i.e., an unknown modification.

98. The next Examples subsection is titled “Detection of Glucose-Modified 5-hydroxymethylcytosine.” EX1007, 79:19-46. In this experiment, DNA templates were modified with glucose-modified 5-hydroxymethylcytosine at the positions shown in the triangle of FIG. 25. “The two templates are identical except for the glucose-modified hydroxymethylcytosines in the modified template. . . .”⁴ EX1007, 79:41-43. Because the glucose-modified 5-hydroxymethylcytosine experiments used identical DNA templates, the nucleotide identity of the template sequence is a constant, not a variable. FIG. 25 shows “the ratio of the average IPD in the glucose-modified template to the average IPD in the unmodified template, plotted versus DNA template position.” EX1007, 79:38-41. “Average IPD ratios for glucose-modified hydroxymethylcytosine increased by as much as five-fold relative to detection of unmodified hydroxymethylcytosine. . . .” EX1007, 79:43-46.

⁴ Similar to the 5-methylcytosine templates, this example does not disclose what the contrived DNA template sequences are.

99. The glucose-modified 5-hydroxymethylcytosine experiments of the '338 patent are simple idealized experiments using DNA sequences with contrived patterns of modified bases and unmodified bases along with control sequences having the same primary sequence but without any modification. IPD was measured at a known modification site and surrounding bases of a DNA template and then compared to a modified DNA template. Those experiments showed a correlation between IPD ratios and a glucose-modified 5-hydroxymethylcytosine modification at the site of the modification and at bases upstream and downstream of the modification. However, the experiments did not show any *de novo* detection of DNA methylation, i.e., an unknown modified base.

100. The next Examples subsection is titled "Detection of Modified Bases in RNA Templates." EX1007, 79:47-80:35. In this experiment "RNA templates containing N⁶-methyladenosine modifications at defined positions were produced by solid-phase nucleic acid synthesis. In addition, control RNA templates with identical sequences were synthesized without the above-mentioned modification." EX1007, 79:52-80:1. Because these experiments used identical RNA templates, the nucleotide identity of the template sequence is a constant, not a variable. "[T]he values for IPD are sensitive to the presence of N⁶-methyladenosine modification, e.g., the distances between individual pulses increase in the presence of this modification." EX1007, 80:31-34.

101. The N⁶-methyladenosine modified RNA experiments of the '338 patent are simple idealized experiments using RNA sequences with contrived patterns of methylated and unmethylated bases along with control sequences having the same primary sequence but without any methylation. IPD was measured between a known methylated template and an unmethylated template. Those experiments showed a correlation between IPD and a N⁶-methyladenosine modification. However, the experiments did not show any *de novo* detection of RNA methylation, i.e., an unknown methylated base.

102. In sum, all the experiments in the Examples sections used contrived nucleic acid templates, comparing template(s) with known modification(s) with an unmodified control template having an identical primary sequence. None of the experiments showed *de novo* detection of nucleic acid modification, e.g., one or more unknown modified base(s), and unknown modified ribose. All of the experiments measured IPD. Only two experiments measured PW ratios which exhibited substantially lower signal-to-noise ratio than the IPD ratios of the same templates. Based upon the experiments and results of the '338 patent, a person of ordinary skill in the art would conclude that IPD exhibits substantially higher signal-to-noise ratio and thus much greater sensitivity to a modified base than pulse width. With respect to nucleotide identity, none of the examples used or suggested nucleotide identity as a feature for an input data structure. When sequence context

was mentioned in the disclosure of the Examples section, sequence context was used to refer to the bases around the modified base, *i.e.*, a window of bases, not nucleotide identities as input data. And, like Flusberg, when reviewing polymerase kinetics over a sequence context, ratios of IPD and PW were used as a normalization that removes the contribution of nucleotide identity. Consequently, the identity of the nucleotides of the template sequences is a constant in the experiments that observed polymerase kinetics over a sequence context, not a variable.

103. In my opinion, the '338 patent does not teach the input data structure of the '794 patent. I have reviewed Dr. Dessimoz's citations from the '338 patent that support his opinion that the '338 patent discloses the features of the input data structure of claim 1. EX1002, pp. 58-72, ¶¶101-20.

104. Dr. Dessimoz cites the '338 patent (EX1007), 66:46-67:4 for the proposition that the '338 patent “teaches ‘development of a classifier for distinguishing methylated from unmethylated cytosines in a nucleic acid template,’ for which a ‘set of features is measured for every pulse,’ including ‘pulse width’ and ‘[inter]pulse duration’ for each nucleotide.” EX102, p. 59, ¶103 (also cited on p. 60, ¶104). The cited section of the '338 patent states:

In certain aspects, the invention provides a general-purpose approach to discriminating between Me⁺ and Me⁻ using features in a real-time sequencing-by-synthesis trace comprising signals emitted during the incorporation of optically detectable nucleotides into a nascent strand

by a polymerase enzyme. . . . 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

$$\vec{f} = \{f_1^{i-3}, f_2^{i-3}, \dots, f_k^{i-1}, f_1^i, f_2^i, \dots, f_k^i, \dots, f_1^{i+3}, f_2^{i+3}, \dots, f_k^{i+3}\}.$$

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. EX1007, 66:46-67:4.

105. This does not disclose the input data structure of claim 1. First, nucleotide identity as part of an input data structure is not disclosed or suggested to a person of ordinary skill in the art. The reference to local context does not refer to nucleotide identity, but rather the window around a cytosine that the pulse values should be called. Second, the disclosure does not teach a person of ordinary skill in the art to use IPD and PW together in an input data structure, but lists them in a set of features that might be used along with pulse height and pulse amplitude

variability. As I explained in paragraph 26, feature selection is unpredictable. To select appropriate features, the predictive power of the features needs to be assessed, which is missing from the '338 patent disclosure.

106. Dr. Dessimoz addresses “the identity of each nucleotide” in paragraph 111 of his declaration. EX1002, p. 66, ¶111. Dr. Dessimoz states that “[i]ncluding the identity of each nucleotide in the data structure, as taught by PacBio-BAM, is consistent with the Flusberg Publications’ teachings.” Dr. Dessimoz cites the '338 patent (EX1007), 62:28-39 as “the local sequence context of the template strand can also influence and inhibit the activity of the polymerase along the template.” What the patent states is:

Other features that are related to pausing and can contribute to a full model of the phenomenon. In particular, the local sequence context of the template strand can also influence and inhibit the activity of the polymerase along the template. For example, the local sequence context that influences and/or inhibits activity of the polymerase may extend for at least about one, two, three, four, five, seven, ten, fifteen, twenty nucleotide positions, and these positions may lie upstream or downstream of the modification, or may flank the modification in the template. EX1007, 62:28-39.

107. The use of the term “local sequence context” in the above passage does not refer to nucleotide identity, but rather the window of nucleotides around a modification. The polymerase interacts with DNA over an overextended window of

bases, which influences the reaction kinetics, e.g., IPD ratios. This passage does not teach or suggest using nucleotide identity in an input data structure.

108. Dr. Dessimoz next cites the '338 patent (EX1007), 64:53-55 which states: "It [conditional random field vector] can also better integrate knowledge of the inter-pulse spacing and other information, such as sequence context, into a broad model." EX1002, p. 66, ¶111. That disclosure is directed to differentiating between pauses and sequencing states, not determining whether a base is modified or not. EX1007, 64:29-55. Sequence context refers to a window of bases, not the nucleotide identity of the nucleotides in that window.

109. Dr. Dessimoz next cites the '338 patent (EX1007), 66:4-45 to support the parenthetical "probability of methylation related to 'neighboring sequence context' and features such as IPD." EX1002, p. 66, ¶111. That passage is focused on IPD – "In particular interpulse duration (IPD) information from multiple positions can be used to determine whether or not a given cytosine is methylated. . . ." EX1007, 66:7-9. The neighboring sequence context reference is as follows:

The individual weights for the multiple positions (all together which would likely sum to 1) could be based on a combination of the following metrics, assuming we are comparing two templates that are identical aside for one being methylated and the other being unmethylated at a given position: the ratio of or difference in IPD between the two templates at that given position; the statistical significance of the distinguishability between the IPD distributions of

the two templates at that given positions; the number of observations used when creating the IPD distributions; and the neighboring sequence context. EX1007, 66:26-36.

110. As the above passage explains, the two templates being compared are identical which means the nucleotide identity is a constant, not a variable. Consequently, the use of the term “neighboring sequence context” in the above passage does not refer to nucleotide identity, but rather the window around a modification.

111. Dr. Dessimoz next quotes the '338 patent (EX1007) at 5:33-46 in a parenthetical “a combination of sequence context and methylation status affected by interpulse duration.” EX1002, p. 66, ¶111. That quotation is not found at 5:33-46. That citation does not refer to sequence context at all. The parenthetical quote appears to be taken from the Examples section of the 5-methylcytosine experiments.

Methylation sequencing on a SMRTTM Sequencing platform (see, e.g., P. M. Lundquist, et al., Supra) was performed on short, synthetic DNA oligos with contrived patterns of methylated and unmethylated bases, along with control sequences having the same primary sequence but without any methylation. These templates provided unequivocal fluorescence pulse patterns and tempos that demonstrated how the combination of sequence context and methylation status affected interpulse duration. For example, SMRTTM sequencing experiments were performed using synthetic DNA templates that only differed by a single methylated vs. unmethylated cytosine. The difference in average

interpulse durations between the two templates was visible both at the 5-MeC position and in the vicinity of the 5-MeC position. EX1007, 75:33-46.

112. Because the DNA templates had an identical sequence, nucleotide identity is a constant not a variable. “Sequence context” in the above passage does not refer to nucleotide identity, but rather the window around a modification.

113. Dr. Dessimoz next cites the '338 patent (EX1007), 64:29-65:20 twice for “prediction based ‘on observed pulse data (both basecalls and other pulse features)’” in a model to predict nucleotide modifications. EX1002, p. 66, ¶111. The disclosure at 64:29-65:20 is directed to a model for differentiating between pauses and sequencing states, not nucleotide modifications.

A more powerful algorithmic architecture for segmentation is the use of the conditional random field framework (CRF). The object is to predict the conditional probability of a signal arising from a pause or sequencing state given the observed pulses:

$$p(y | x) = \frac{\exp[w^T F(x, y)]}{\sum_{y'} \exp[w^T F(x, y')]} ,$$

where y is the sequence of desired labels (pause, sequencing), x is the observed pulse data (both basecalls and other pulse features), w is the weight vector learned from the training data, and the F function is the feature vector. EX1007, 64:29-43.

114. Basecall refers to determining, one base at a time, whether the base is A, C, G, or T, not keeping track of the nucleotide identities of a string of bases. The use of basecalls and other pulse features for predicting the conditional probability of a signal arising from a pause or a sequencing state does not teach a person of ordinary skill in the art to use nucleotide identity in an input data structure to detect nucleic acid modifications.

115. Lastly, Dr. Dessimoz cites the '338 patent (EX1007) at 7:4-17 as a “pause indicative of modification.” EX1002, p. 66, ¶111. But the disclosure at 7:4-17 is directed to “methods for bypassing one or more modifications in a single nucleic acid template during a sequencing-by-synthesis reaction.” EX1007, 7:4-6. The disclosure is uninformative to a person of ordinary skill in the art with respect to using nucleotide identity as input data for modification detection.

116. In sum, the disclosure of the '338 patent does not teach or suggest nucleotide identity as a feature for an input data structure. The '338 patent uses the term “sequence context” or “local context” to refer to a window of bases surrounding a putative base modification. When the parenthetical quotations and citations from Dr. Dessimoz’s declaration are closely examined, they reflect that “sequence context” and “local context” refer to a window of bases surrounding a putative base modification.

117. With respect to pulse width, Dr. Dessimoz cites the '338 patent (EX1007), 66:46-67:4. EX1002, p. 71, ¶118. As I have explained above in paragraph 104, this disclosure does not teach a person of ordinary skill in the art to use IPD and PW together in an input data structure, but lists PW and IPD in a set of other features that might be used. Dr. Dessimoz also cites disclosure that nucleic acid modifications can alter IPD or PW. EX1002, p. 71, ¶118, EX1007, 9:41-46, 22:49-54, FIGs. 12-13 and 15-19.

The change in processing can occur at the modification or further modification, or can occur at one or more positions upstream or downstream of the modification or further modification, and can be a kinetic change such as an alteration in interpulse duration or pulse width. EX1007, 9:41-46.

For example, the presence of modified bases in a template nucleic acid has been shown to change the width of a pulse and/or the interpulse duration (IPD), either at the modified base or at one or more positions proximal to it. A change in pulse width may or may not be accompanied by a change in IPD.

EX1007, 22:49-54.

118. The above passages disclose that modified bases affect IPD and/or PW. It does not teach to use PW and IPD together to detect base modifications.

119. Lastly, Dr. Dessimoz cites FIGs. 12-13 and 15-19. EX1002, p. 71 ¶118. FIGs. 12, 13, 15, 16, and 18 are directed to IPD data and FIGs. 17 and 19 address PW data. Because the signal-to-noise ratio, defined by the dynamic range around a

methyated position compared to the dynamic range in background positions, of the PW ratios is so small compared to the IPD (paragraphs 91 and 94 above) and because the focus of the experimental data was on the IPD ratios, a person of ordinary skill in the art would conclude based upon the experimental data disclosed in the '338 patent that IPD was the input data for determining nucleic acid modifications, not PW.

120. To be complete, Dr. Dessimoz also cites the '338 patent as supporting the use of PW in paragraphs 79, 80 and 83 of his declaration. Paragraph 79 cites the '338 patent (EX1007) at 23:14-22; 64:29-42; 66:46-67:4 and FIGs. 1 & 2. EX1002, p. 42, ¶79. As I explain above at paragraph 108, the disclosure at 64:29-42 is directed to a model for differentiating between pauses and sequencing states, not nucleotide modifications. As I explain above at paragraph 104, the disclosure at 66:46-67:4 does not teach a person of ordinary skill in the art to use IPD and PW together in an input data structure, but lists PW and IPD in a set of other features that might be used. FIG. 1 is an exemplary illustration of SMRT sequencing. FIG. 2 is explained at 23:14-22 and merely shows a methyated base affects IPD and PW:

FIG. 2 provides illustrative examples of various types of reaction data in the context of a pulse trace including IPD, pulse width (PW), pulse height (PH), and context. FIG. 2A illustrates these reaction data on a pulse trace generated on an unmodified template, and FIG. 2B illustrates how the presence of a modification (5-MeC) can elicit a

change in one of these reaction data (IPD) to generate a signal (increased IPD) indicative of the presence of the modification. EX1007, 23:14-22.

121. Paragraph 80 of Dr. Dessimoz's Declaration cites the '338 patent (EX1007), 66:46-67:36, 66:46-68:48. EX1002, pp. 42-43, ¶80. As I explain above at paragraph 104, the disclosure at 66:46-67:4 (EX1007) does not teach a person of ordinary skill in the art to use IPD and PW together in an input data structure, but merely lists PW and IPD in a set of features that might be used. Similarly, the disclosure from 67:4-68:48 (EX1007) does not teach a person of ordinary skill in the art to use IPD and PW together in an input data structure, but lists PW in a set of other features that might be used: "the approach can be extended to pulse metrics other than IPD, such as pulse width, branch rate, mismatch rate, deletion rate, etc." EX1007, 67:65-67.

122. Paragraph 83 of Dr. Dessimoz's Declaration cites the '338 patent (EX1007), 4:66-5:3, 22:45-54, 78:18-43, 79:2-5, FIGs. 18-19. EX1002, p. 44, ¶83. The disclosure at 78:18-43 (EX1007) is directed to the 8-oxoguanosine experiments I discussed above in paragraphs 93-95. The disclosure at 79:2-5 (EX1007) is addressed above at paragraphs 96-97, directed to experiments that added RNA ribonucleotides to DNA templates and measured the ratio of the average IPD. FIGs.

18 and 19 are directed to IPD data. That leaves the citations to 4:66-5:3, 22:45-54 (EX1007). As shown below 4:66-5:3 says nothing about PW.

In some embodiments, a change occurs only at the modification, and in other embodiments the change occurs at one or more positions proximal to the modification, which may also include the modification position. EX1007, 4:66-5:3.

And, EX1007, 22:45-54, states that modifications may alter PW.

For example, the presence of certain modifications alters not only the processing of the template at the site of the modification, but also the processing of the template upstream and/or downstream of the modification. For example, the presence of modified bases in a template nucleic acid has been shown to change the width of a pulse and/or the interpulse duration (IPD), either at the modified base or at one or more positions proximal to it. A change in pulse width may or may not be accompanied by a change in IPD. EX1007, 22:45-54.

123. In sum, the '338 patent does not disclose using IPD and PW as a combination for an input data structure to detect base modification. The '338 patent includes IPD and PW in a listing of possible features including other features such as pulse height, pulse amplitude variability in one list and branch rate, mismatch rate, deletion rate in another. Because the '338 patent teaches that IPD ratio exhibits substantially higher signal-to-noise and thus much greater sensitivity to a modified base than PW, a person of ordinary skill in the art would be discouraged from

following the path of using PW as input data to detect base modification and would be led in the direction of IPD ratios which is divergent from the path followed by and claimed in the patent – the combination of nucleotide identity, IPD and PW. *See* Schadt (EX1010), where the '338 patent joint-inventors Jonas Korlach and Tyson A. Clark mentioned: “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 IPD.” Schadt (EX1010), p. 130.

124. There is no teaching or suggestion to a person of ordinary skill in the art in the '338 patent to use nucleotide identity as input data, either. The '338 patent teaches using IPD ratios over a control template. As the '338 patent joint-inventors Jonas Korlach and Stephen W. Turner explain in Suzuki (EX2005), IPD ratios “eliminate the context-dependent fluctuation of the IPD values.” EX2005, p. 2912. A person of ordinary skill in the art would be discouraged from following the path of using nucleotide identity as input data to detect base modification because the '338 patent teaches using IPD ratios as a normalization to remove the contribution of nucleotide identity. A person of ordinary skill in the art would be led in the direction of IPD ratios which is divergent from the path followed by and claimed in the patent - the combination of nucleotide identity, IPD and PW.

X. The PacBio BAM does not teach or suggest to a person of ordinary skill in the art the “input data structure” of claim 1.

125. BAM stands for a Binary Alignment/Map (BAM). The BAM file has an industry standard format for storing the comprehensive raw data of genome sequencing; it consists of compressed binary representation of the Sequence Alignment Map (SAM)-files. The seminal paper for the BAM file format is Li et al., “*The Sequence Alignment/Map Format and SAMtools*,” *Bioinformatics* 25(16) (2009), pp. 2078-2079 (EX2004).

126. The PacBio BAM “format is a binary compressed, record-oriented container format for raw or aligned sequence reads.” EX1011, p. 1. The PacBio BAM file contains both sequence and kinetics data. As set forth in EX1011, PacBio stores different types of data in the BAM format (e.g., “ZMW reads from movie,” “Analysis-read subreads from movie,” etc.). EX1011, p. 2. The PacBio BAM file is not “a preexisting, off-the-shelf option” as Dr. Dessimoz opines (EX1002, p. 44, ¶82) for an “input data structure,” because the input data must be parsed out from the extraneous information in the PacBio BAM file and assembled into an input data structure, e.g., a matrix of same sized windows to train the model.

127. Because the PacBio BAM is compressed, it cannot be directly read by a trainable model, it must be decompressed first. A decompressed BAM file cannot be the input data structure of claim 1 because it comprises text data, not an input data array of numerical values as required for statistical analysis with a trainable model.

The PacBio BAM file includes extraneous information both numeric and non-numeric that needs to be parsed out before the data could be used by a trainable model such as SAM Flags, subread name, context flag, number of passes, base quality, position end, position start, read quality, signal-to-noise, zmw number and read group. EX2012 to EX2015. Even after all the extraneous information is removed from a PacBio BAM file, the data needs to be structured into same sized windows for training and testing.

128. A PacBio chip has thousands to millions of wells in which thousands to millions of nucleic acid molecules are sequenced. The latest PacBio high density chip has 25 million ZMWs. EX2027, p. 1. In 2011, the PacBio chip had 45,000-75,000 wells. EX1026, p. 2. Whether the chip has millions or thousands of wells, DNA sequences of various sizes are sequenced and stored in the PacBio BAM file. The size of each sequence in each well is random. Some sequence lengths may be the exact same length, but many, if not most, will be different lengths. The PacBio BAM file includes all the sequence data from each PacBio chip. That PacBio BAM file will have thousands to millions of sequence reads based upon the number of wells in the chip. While the '794 patent discloses that the "measurement windows may be the entire template molecule," (EX1001, 34:46-47), those "measurement windows" must be the same size to train the model and then detect base modification. Consequently, the measurement window cannot be the set of all nucleotides in the

PacBio BAM because the sequence read lengths stored in the PacBio BAM file will vary from sequence read to sequence read.

129. While the PacBio BAM notes that the kinetic data may be used for base modification detection, the PacBio BAM file has many other uses including visualization of read alignments in a genome browser (EX2010), peak calling analysis (EX2011), transcript assembly and quantification (EX2007), and calling variants (EX2008) or haplotypes (EX2009). The PacBio BAM's reference that kinetic data can be used for "base modification detection" does not teach or suggest to a person of ordinary skill in the art how to create an input data structure or to include the combination of features of the claimed input data structure. EX1011, p. 1. It just indicates that the PacBio BAM format saves disk space without affecting the kinetic data for "base modification detection."

130. Because the PacBio BAM is the data generated from all SMRT sequencing reactions of a PacBio chip, PacBio BAM does not have a target position. It is just the data. Dr. Dessimoz opines at paragraph 113 that the PacBio BAM teaches the feature of "a position of the nucleotide with respect to a target position within the window." EX1002, ¶113. Dr. Dessimoz cites the PacBio BAM, EX1011, pp. 1-2 which addresses a sequence alignment between SMRT sequence read and a reference genome. EX1011, p. 1 ("Upon alignment, generally only a subsequence of the query will align to the reference genome."). A person of ordinary skill in the

art would understand that the recitation of the word “target” on page 2 of EX1011 (“‘ref.’/‘target’ coord. System”) in a sequence alignment does not disclose or suggest the claimed “target position within the window.” The term “target” in the EX1011, p. 2 refers to the reference sequence, i.e., “target sequence” to which the SMRT sequence read is aligned to generate an aligned BAM file. It is not a target position for base modification detection, and does not teach or suggest creating an “input data structure” having “a position of the nucleotide with respect to a target position within the window” as claimed.

131. In my opinion, the PacBio BAM adds nothing to the teachings of Flusberg and the '338 patent because it is the stored data generated in the normal course of SMRT sequencing. Both Flusberg and the '338 patent disclose that the data used in the disclosed experiments were data generated in the normal course of SMRT sequencing.

XI. A person of ordinary skill in the art would not have been motivated to combine Flusberg, the '338 patent, and the PacBio BAM to arrive at the invention of claim 1.

132. A person of ordinary skill in the art would not have a reasonable expectation of success at achieving claim 1 because the teachings of Flusberg and/or the '338 patent do not teach or suggest the use of nucleotide identity as an input data, let alone the combination of nucleotide identity with PW and IPD as I have explained above. The PacBio BAM file is merely the stored data generated in the normal

course of SMRT sequencing and does not teach or suggest the selection of input features of claim 1.

133. I have reviewed Dr. Dessimoz's declaration regarding the combination of Flusberg, the '338 patent, and the PacBio BAM file. EX1002, pp.43-50, ¶¶82-90. I disagree that a person of ordinary skill in the art would have a reasonable expectation of success at arriving at claim 1 of the '794 patent, in particular, arriving at the claimed input data structure that can train a model and detect modified bases.

134. As I have explained above, the experiments in Flusberg and the '338 patent show a correlation between kinetic features, particularly IPD ratios and nucleic acid modifications, e.g., DNA methylations, but neither reference has experiments that detect DNA modifications *de novo*. Based upon the limited experiments performed by Flusberg and disclosed in the '338 patent, a person of ordinary skill in the art would not have reasonable expectation of success of achieving the modification detection method of claim 1 based upon those teachings individually or as combination with the PacBio BAM because the input data structure is not taught or suggested.

135. To support his opinion that a person of ordinary skill in the art would have "good reason to combine the Flusberg with PacBio-BAM to achieve what is recited in the claims with a reasonable expectation of success," Dr. Dessimoz cites several papers from PacBio researchers and collaborators: Flusberg (2010), EX

1006, Clark (2011), EX1008, Schadt (2013), EX1010, Loomis (2013), EX1009, and Feng (2013), EX1015.

136. Flusberg's co-authors Tyson Clark and Jonas Korlach explain in Schadt (EX1010) the simplicity and shortcomings of Flusberg:

While this simple ratio approach is effective at identifying high confidence KVEs, it does not maximally leverage the available information. For example, the ratio method only makes use of mean kinetic variation behavior at a single site, not taking into account other feature of kinetic parameter distributions, such as number of observations, variance, kinetic information at neighboring bases that may reflect modification at a given test site, and interactions between neighboring sites. EX1010, p. 130.

137. Similar to Flusberg, Clark (EX1008) used the same Flusberg IPD ratio method conducting experiments with DNA templates having contrived patterns of methylated and unmethylated bases along with control sequences having the same primary sequence but without any methylation. IPD was measured at a known methylation site and surrounding bases of a DNA template and then compared to an unmethylated control DNA template. Those experiments showed a strong correlation between IPD ratios and 8-oxoguanine (4X increase), 8-oxoadenine (10X increase), O6-methylguanine, and O4-methylthymine (10-20X increase) modifications at the site of the modifications. EX1008, pp. 2, 5. But, none of the corresponding pulse width ratios showed a strong correlation.

138. The papers that followed, Schadt (2013) (EX1010), Loomis (2013) (EX1009), Feng (2013) (EX1015), and Suzuki (2016) (EX2005) all focused on IPD and did not consider PW in their analysis. “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 IPD.” Schadt (EX1010), p. 130. PacBio’s kineticsTools as described in the PacBio kineticsTools manual (EX1021) uses IPD, not PW. EX1021, p. 1, Overview: “kineticsTool loads IPDs observed at each position in the genome, and compares those IPDs to value expected for unmodified DNA, and outputs the result of this statistical test.”

139. Schadt “developed a model that considers IPD information at sites flanking the site of interest and allows for neighboring sites to interact. . . .” EX1010, p. 132. While Schadt acknowledges the importance of local sequence context, Schadt uses the term “local sequence context” in the same way as Flusberg and the ’338 patent, namely as a window around a putative base modification. “[A] significant proportion of the variance in the IPD measures can be explained by local sequence context. In fact, seven bases upstream of and two bases downstream from the incorporation site explain >80% of the variance in the IPD measures (Fig. 2A).” EX1010, p. 130.

140. Schadt's teaching that over 80% of the variance in IPD measures is found in a nine-base window around a base modification does not teach or suggest using PW or nucleotide identity as input data. Schadt's teaching that PW is not as sensitive to shifts in sequence context and development of an IPD based model that excludes PW as an input teaches away from the use of PW in an input data structure. And, the additional journal articles from PacBio researchers and collaborators – Loomis (2013) (EX1009), Feng (2013) (EX1015), and Suzuki (2016) (EX2005) using an IPD model reinforces that teaching, as does PacBio's kineticsTools as described in the PacBio kineticsTools manual (EX1021). Because of those teachings a person of ordinary skill in the art would be discouraged from following the path of using PW as input data to detect base modification.

141. Lastly, Dr. Dessimoz opines that kinetic parameters and sequence context were used in other contexts, specifically Oxford Nanopore sequencing. As explained in the '794 patent while Oxford Nanopore sequencing allows for long read like SMRT sequencing, the approaches are fundamentally different.

As the principle of single molecule, real-time sequencing (e.g., the Pacific Biosciences SMRT system) is different from that of a non-optical based nanopore system (e.g., by Oxford Nanopore Technologies), approaches for base modification detection developed for such non-optical based nanopore system cannot be used for single molecule, real-time sequencing. For example, a non-optical nanopore system is not designed for capturing the patterns of fluorescent signals

produced by immobilized DNA polymerase based DNA synthesis (employed by single molecule, real-time sequencing such as by the Pacific Bio-sciences SMRT system). As a further example, in the Oxford Nanopore sequencing platform, each measured electric event is associated with a k-mer (e.g., 5-mer) (Q. Liu et al., 2019). However, in the Pacific Biosciences SMRT sequencing platform, each fluorescent event is generally associated with a single incorporated base. Furthermore, a single DNA molecule would be sequenced multiple times in Pacific Biosciences SMRT sequencing including Watson and Crick strands. Conversely, for the Oxford Nanopore long-read sequencing approach, sequence readout is performed once for each of the Watson and Crick strands. EX1001, 17:42-63.

142. These different sequencing approaches result in different features. For example, the '338 patent discloses interpulse duration (IPD), pulse width (PW), pulse height (PH), local signal-to-noise ratio, dye channel, the χ^2 metric for the measured spectrum, and pulse amplitude variability as possible features (EX1007, 23:14-17, 63:19-22, 66:58-60) whereas Liu compared a 7 feature description of an event to a 57 feature description of an event which have no apparent overlap to the '338 patent features. EX1022, pp. 3-4. The approaches are simply not comparable. The experimental journey of Liu's work also illustrates that feature selection is an important process in the use of a trainable model. Merely having a data repository, like the BAM file, does not help.

XII. Modification detection using the claimed input data structure provided unexpected results compared to Flusberg and the '338 patent.

143. Based upon Flusberg and the '338 patent, a person of ordinary skill in the art would expect that IPD would be far more predictive of base modification than PW. An unexpected result of the '794 patent is the predictive power of PW as a feature which had a 0.85 AUC compared to 0.73 AUC for IPD. Thus, the strong predictive power of the combination of features that make up the input data structure of claim 1, nucleotide identity and relative position to the target position ("Seq"), IPD, and PW with an 0.94 AUC is unpredicted. EX1001, 32:18-27. FIG. 18 illustrates the results.

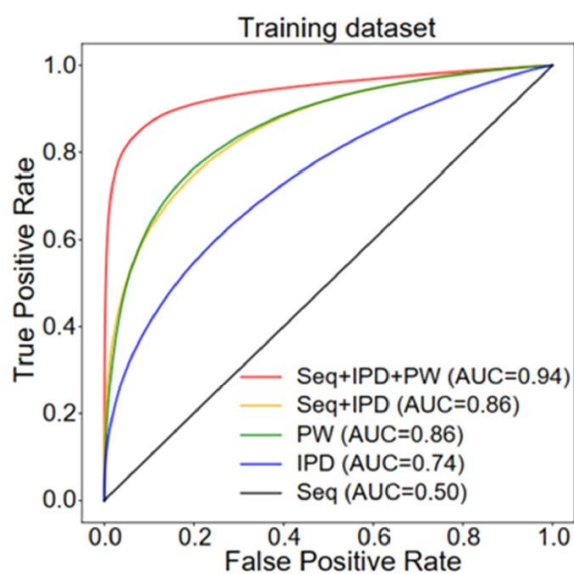


FIG. 18A

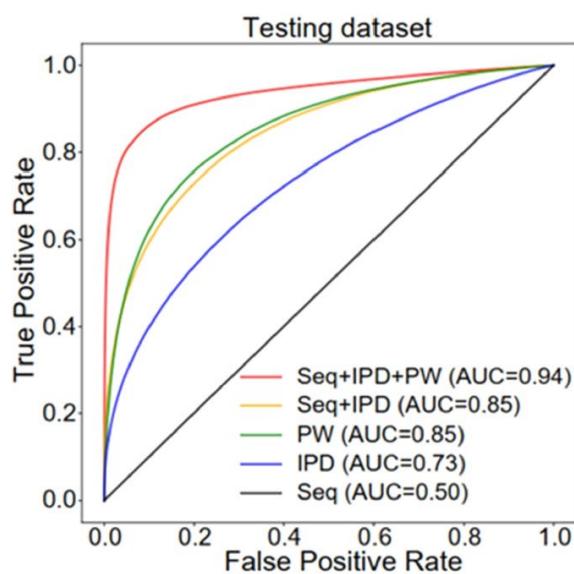


FIG. 18B

EX2001, FIG. 18.

144. For all of the above reasons, it is my opinion that claim 1 of the '794 patent is non-obvious over Flusberg, the '338 patent, and the PacBio BAM. My opinion is the same if Flusberg or the '338 patent is considered alone or together with the PacBio BAM.

I hereby declare that all the statements made in this Declaration are of my own knowledge and true; that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under 18 U.S.C. § 1001 and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

Dated: January 26, 2024

By: 
Peter K. Koo, Ph.D.