

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 2041:
SUPPLEMENTAL 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. '794 patent, Claim 1	16
	C. '794 Patent Specification	18
	1. “input data structure”	18
	2. The '794 patent evaluates the predictive power of features.	24
	3. The '794 patent discloses a methylation detection method correlating well with bisulfite sequencing.	27
	D. Prosecution History	34
VII.	TSE’S HOLISTIC KINETIC MODEL IS THE MODEL OF CLAIM 1	35
VIII.	CLAIM 1 IS NOT OBVIOUS OVER FLUSBERG, THE '338 PATENT, AND THE PACBIO-BAM	42
	A. Flusberg does not Disclose or Suggest to a Person of Ordinary Skill in the Art the “Input Data Structure” of Claim 1.	42
	B. The '338 Patent does not Disclose or Suggest to a Person of Ordinary Skill in the Art the “Input Data Structure” of Claim 1.	50
	C. The PacBio-BAM does not Teach or Suggest to a Person of Ordinary Skill in the Art the “Input Data Structure” of Claim 1.	59
	D. A Person of Ordinary Skill in the Art Would not be Motivated to Combine Flusberg, the '338 Patent, and the PacBio-BAM to Arrive at the Invention of Claim 1.	61

IX.	OBJECTIVE INDICIA OF NON OBVIOUSNESS.....	70
A.	The Method of Claim 1 Solved the Long-Felt Unsolved Need for Direct Detection of DNA Modifications by SMRT Sequencing Without Bisulfite Treatment	70
B.	Unexpected Results.....	76

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 on 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 not obvious over Flusberg, the '338 patent, and 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 2052 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 provided as EX2032 and EX2055.

10. I have previously provided a declaration in this matter that I refer to herein as EX2002.

B. Relevant Legal Principles

11. 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.

12. 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.

13. 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.

14. 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 claim 1 of the '794 patent is obvious in light of Flusberg, the '338 patent and the PacBio-BAM.

15. 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.

16. 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.

17. 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.

18. 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.

19. 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.

20. 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

21. 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

22. 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.

23. In rendering this opinion I have adopted the qualifications of a POSA as set forth in Dr. Dessimoz’s declaration at paragraph 75. EX1002, ¶75.

VI. The ’794 Patent

A. Background

24. 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.

25. 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 identity of the nucleotide,

a position of the nucleotide within the sample nucleic acid molecule,

a width of the pulse corresponding to the nucleotide, and

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

(b) creating an input data structure, the input data structure comprising a window of the nucleotides sequenced in the sample nucleic acid molecule, wherein the input data structure includes, for each nucleotide within the window, the properties:

the identity of the nucleotide,

a position of the nucleotide with respect to a target position within the window,

the width of the pulse corresponding to the nucleotide, and

the interpulse duration;

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

receiving a first plurality of first data structures, each first data structure of the first plurality of 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.

26. Claim 1 of the '794 patent generally relates to a novel combination of properties obtained from optical sequencing signals to detect nucleotide modifications with a trained model. Specifically, those properties are structured in an input data structure which includes a window and several 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.

27. The selection of features for an input data structure is a fundamental problem in model training. It is often unpredictable how a certain feature or combination of features would perform in model training for predictive analysis. For example, features that carry redundant information or no predictive information may introduce excess noise or cause more challenges than benefits in model training. The selection of the appropriate feature or combination of features as data inputs for a trainable model, also referred to as feature engineering, feature construction, feature selection, or feature extraction, is often 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 at 15.

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

29. Because Ground 1 is premised on Flusberg. I address the '794 patent disclosure directed to Flusberg.

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

32. 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.

33. Claim 1 requires that identity of each nucleotide be included as a property in an “input data structure.” Flusberg does not include nucleotide identity as a property in an “input data structure” as recited in claim 1. In contrast, as I

explain below, Flusberg discloses methods that control for nucleotide identity by keeping it constant.

34. The '794 patent further 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. The '794 patent showed PW had a stronger predictive power than IPD for 5mC detection. See the testing dataset (FIG. 18B) by 5mC 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. EX2001, FIG. 18B.

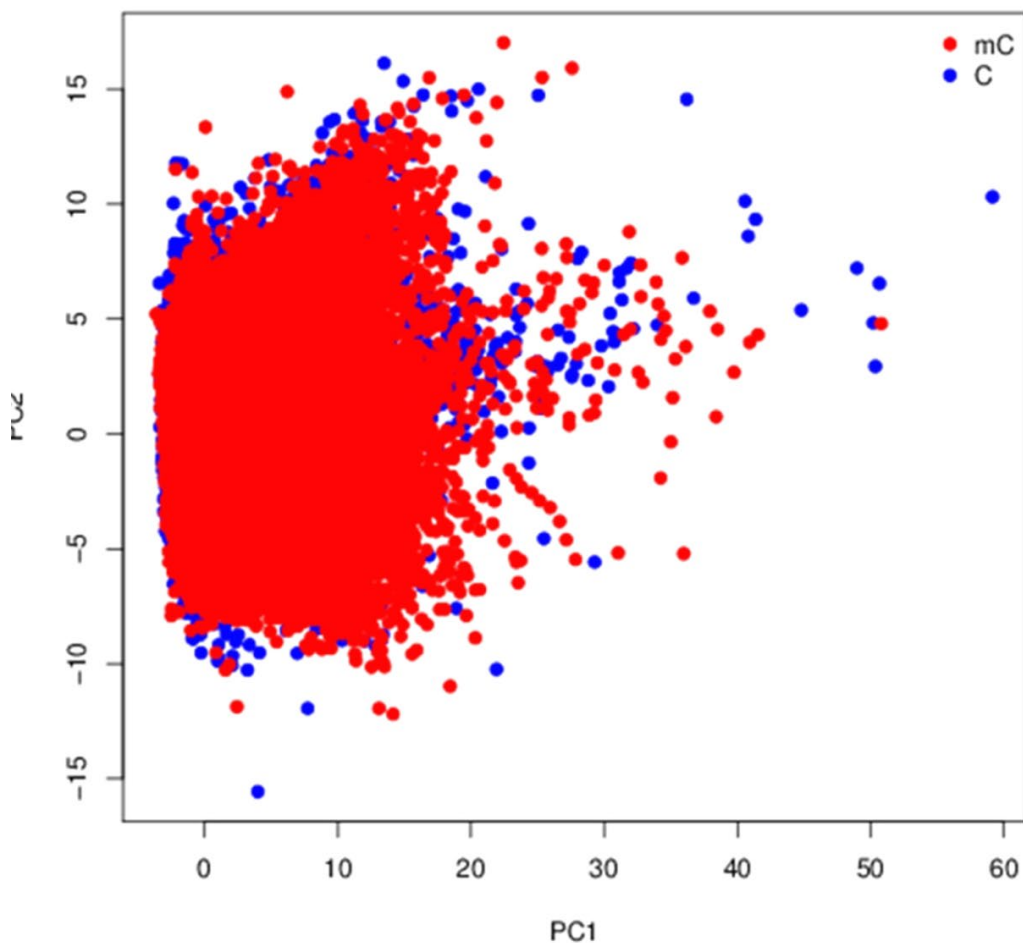
35. In contrast, both Flusberg and the '338 patent showed PW is not as sensitive to base modification (e.g., 5mC and 5hmC) as IPD. See, e.g., EX2002 ¶¶65-66 (Flusberg shows IPD ratios of a known template having known 5mC modifications at known positions (EX1006, Fig. 2b) “had a substantially higher signal-to-noise ratio (2-3X) defined by the ratio of the dynamic range around a modified base compared to background positions” than PW ratio (EX1006, Supplementary Fig. 1a)); EX2002 ¶¶90-91 (The '338 patent discloses IPD ratios of a known template having known 5hmC modifications at known positions (EX1007, FIG. 16) “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” than PW ratio (EX1007, FIG. 17)); and 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 at 130.

36. The '794 patent further tested the methods disclosed 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 downstream 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.

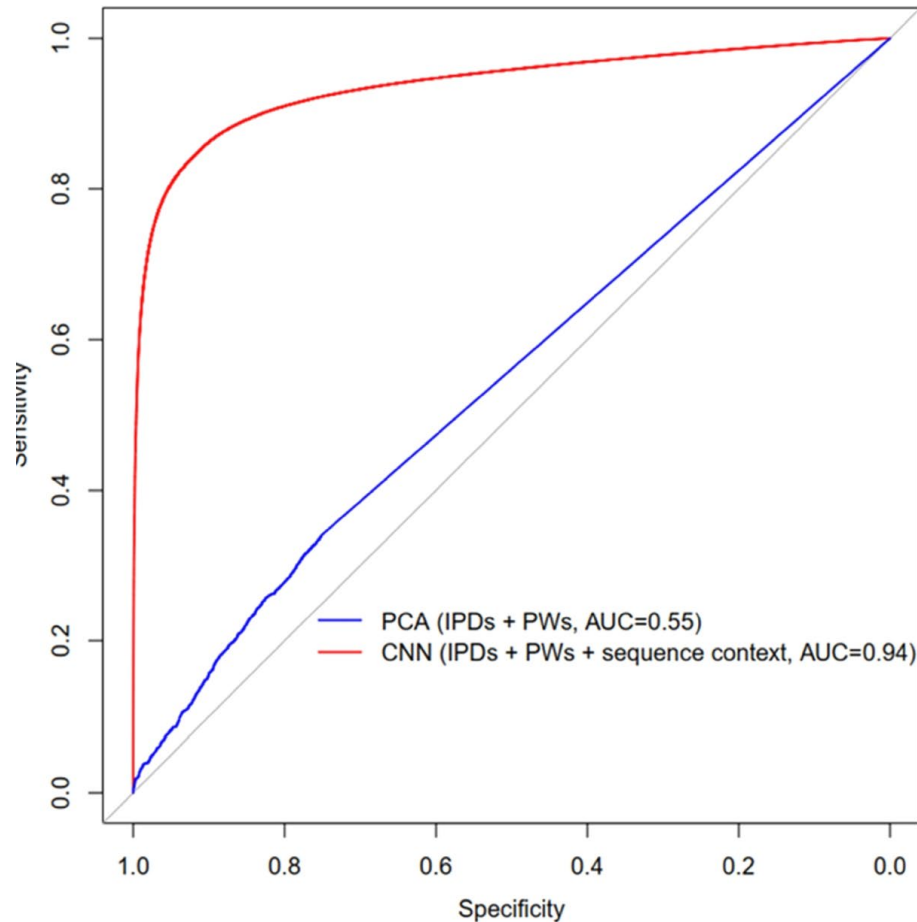
37. As shown above in FIG. 28 from the '794 patent, the principal component analysis of Flusberg was unable to meaningfully separate 5-methylcytosine from unmethylated cytosine.

38. 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.

39. 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 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. '794 patent, Claim 1

40. 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.

41. Claim 1 requires “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 [several] properties.”

42. Claim 1 further requires “creating an input data structure” where the input data structure includes “a window of the nucleotides sequenced in the sample nucleic acid molecule.” And, for “each nucleotide within the window,” claim 1 further requires that the input data structure include the combination of “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.”

43. Claim 1 also requires “inputting the input data structure into a model” where the model is trained by a particular method, including “receiving a first plurality of first data structures.”

44. Claim 1 further recites that “each first data structure comprising values for the same properties as the input data structure.” Thus, in accordance with the plain language of claim 1, a POSA 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 used 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, ¶130.

45. Claim 1 further recites “determining, using the model, whether the modification is present in a nucleotide at the target position within the window in the input data structure.”

C. '794 Patent Specification

1. “input data structure”

46. The '794 patent discloses embodiments of an input data structure to train a model and explained the content of a data matrix as an example of the input data, as I explained in EX2002 ¶¶39-42.

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.

47. 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	408
A	0	0	1.0	0.3	0	0	0	0	1.1	0.8	0	0	0	0	412
C	0	0	0	0	0	0	0	0	0	0	2.3	0.8	0	0	416
G	1.5	1.7	0	0	0	0	0.9	1.2	0	0	0	0	0	0	420
T	0	0	0	0	1.2	0.2	0	0	0	0	0	0	1.5	1.7	

EX2001, FIG. 4 (partial).

48. 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 of 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.

49. 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, EX1001, 23:52-25:41. 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.

50. As further disclosed in the '794 patent and explained in EX2002 ¶44, the data matrices were used to train a model. The data matrix can be generated from only the Watson or the Crick strand, or from a combined data matrix of data generated from both the Watson and Crick strands.

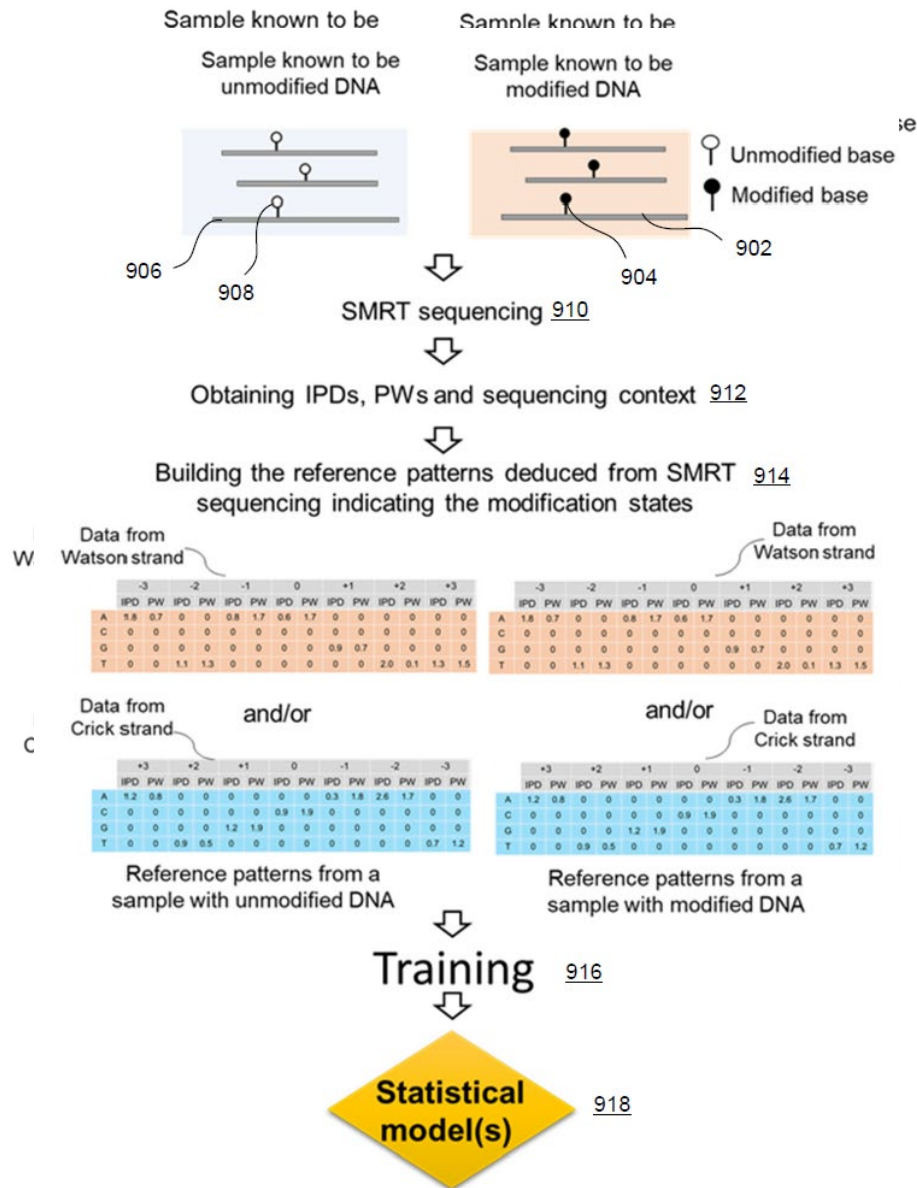
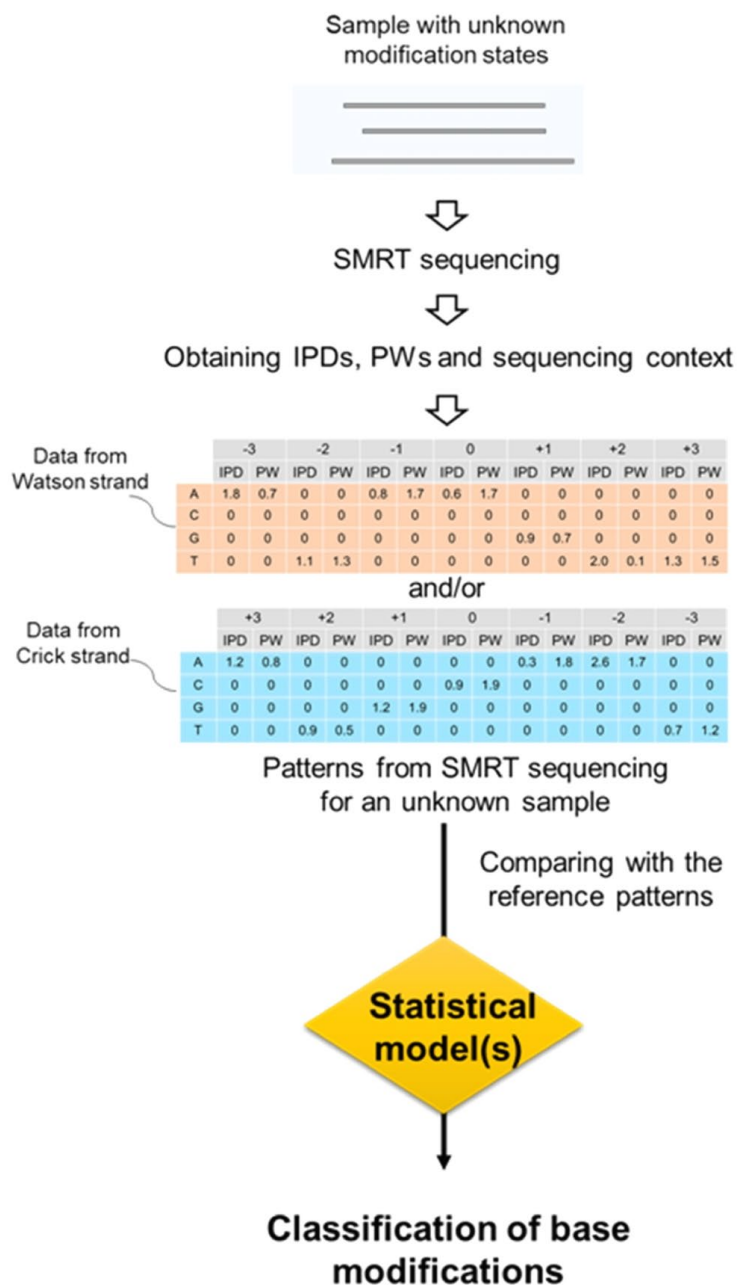


FIG. 9

EX2001, FIG. 9.

51. 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 data structure comprised of 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 of the '794 patent is shown below.



EX2001, FIG. 10.

52. The '794 patent's 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, EX1001, 27:26-30:59. Additional input structures (e.g., various 3-D, 2-D and 1-D structures) are disclosed in the '794 patent, EX1001, 29:54-30:18.

53. 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.

54. 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).

2. The '794 patent evaluates the predictive power of features.

55. The '794 patent “evaluated 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 0.50, 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.

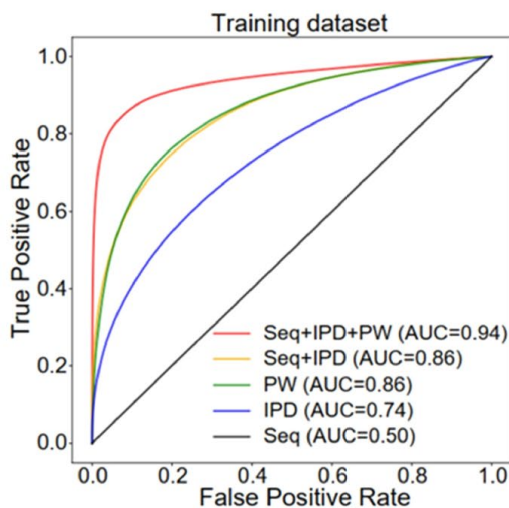


FIG. 18A

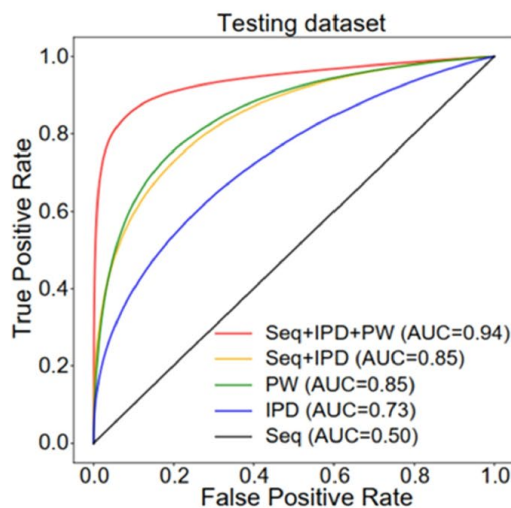


FIG. 18B

EX2001, FIG. 18.

56. 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.

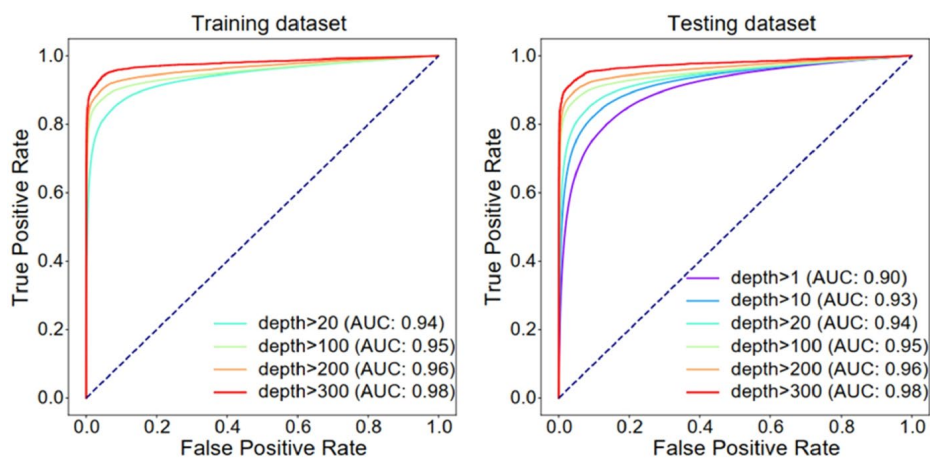


FIG. 19A

FIG. 19B

EX2001, FIG. 19.

57. 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.

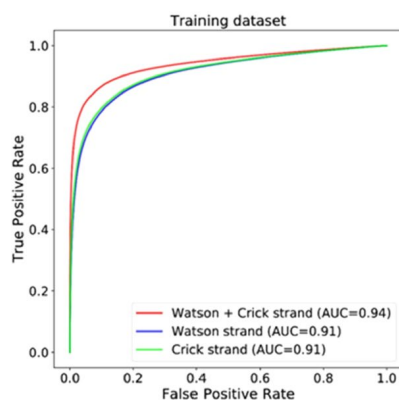


FIG. 20A

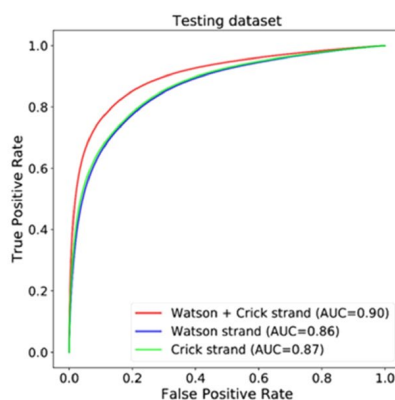


FIG. 20B

EX2001, FIG. 20.

58. The '794 patent also discloses evaluating the power of different numbers of nucleotides upstream and/or downstream of a CpG site in predicting the methylation state of a CpG site, as summarized in EX2002 ¶¶52-55.

3. The '794 patent discloses a methylation detection method correlating well with bisulfite sequencing.

59. The '794 patent explains a need in the art to detect DNA modifications without the use of chemical processing, e.g. bisulfite sequencing as follows:

One disadvantage of such bisulfite-based approaches is that the bisulfite conversion step has been reported to significantly degrade the majority of the treated DNA (Grunau, 2001). Another disadvantage is that the bisulfite conversion step would create strong CG biases (Olova et al., 2018), resulting in the reduction of signal-to-noise ratios typically for DNA mixtures with heterogeneous methylation states. Furthermore, bisulfite sequencing would not be able to sequence long DNA molecules because of the degradation of DNA during bisulfite treatment. Thus, there is a need to determine the modification of bases of nucleic acids, without prior chemical (e.g. bisulfite conversion) and nucleic acid amplification (e.g. using the PCR).

EX1001, 2:19-31.

60. Bisulfite sequencing treats DNA samples with bisulfite to convert unmethylated cytosine to uracil, meanwhile the methylated cytosine remains unchanged. The treated DNA is then PCR- amplified which converts all the uracils into thymines. The amplified DNA can then be analyzed by microarray-based techniques or by 2nd generation DNA sequencing technologies, e.g., Illumina sequencing. The '338 patent, EX1007 at 2:28-39. By the time the '794 patent was

filed, a variety of bisulfite procedures had been developed. *See, e.g.*, Birknerova et al., EX2039 at 15, Beaulaurier et al., EX2038 at 158, and Gouil et al., EX 2040 at 640.

61. From the time of the Flusberg publications¹ through the filing of the '794 patent, bisulfite sequencing was the most common method to detect methylated cytosine and widely recognized as “the gold standard” for single-nucleotide resolution methylation profiling and quantitative DNA methylation analysis, especially for characterizing 5mC. The '338 patent (2009), EX1007 at 2:23-24 (“Bisulfite sequencing is the current method of choice for single-nucleotide resolution methylation profiling. . . .”), Flusberg (2010), EX1006 at 461 (“Currently, the most common technique for studying cytosine methylation involves bisulfite treatment”); PacBio White Paper (2012), EX1016 at 1 (“Most high-throughput techniques focus only on cytosine methylation and involve both bisulfite sequencing to convert unmethylated cytosine nucleotides to uracil nucleotides. . . .”); Beaulaurier et al. (2019), EX2038 at 158 (“Despite its reputation as the gold standard for characterizing 5mC in eukaryotic genomes (owing to its high sensitivity, accuracy and compatibility with NGS technologies)”), the '794 patent (2019),

¹ “Flusberg publications” refers collectively to the Flusberg EX1006 and the '338 patent EX1007.

EX1001, 2:5-6 (“One widely used method to measure DNA methylation is through the use of bisulfite sequencing (BS-seq) . . .”), Tse (2021), EX1030 at 1 (“The most widely used method for detecting 5mC involves bisulfite treatment, followed by methods such as PCR, or massively parallel DNA sequencing (i.e., bisulfite sequencing [BS-seq]. . .”), Sang (2021), EX2042 at 441 (“[B]isulfite sequencing is widely used to determine the patterns of genomic methylation as a gold standard technology”); Birknerova et al. (2022), EX2039 at 15 (“Whole-genome bisulfite sequencing (WGBS) is currently the gold standard DNA methylation profiling technology. . ..”); Wang et al. (2023), EX2043 at 10 (“BS [bisulfite] sequencing and its derivative methods are the gold standards for quantitative DNA methylation analysis. . ..”).

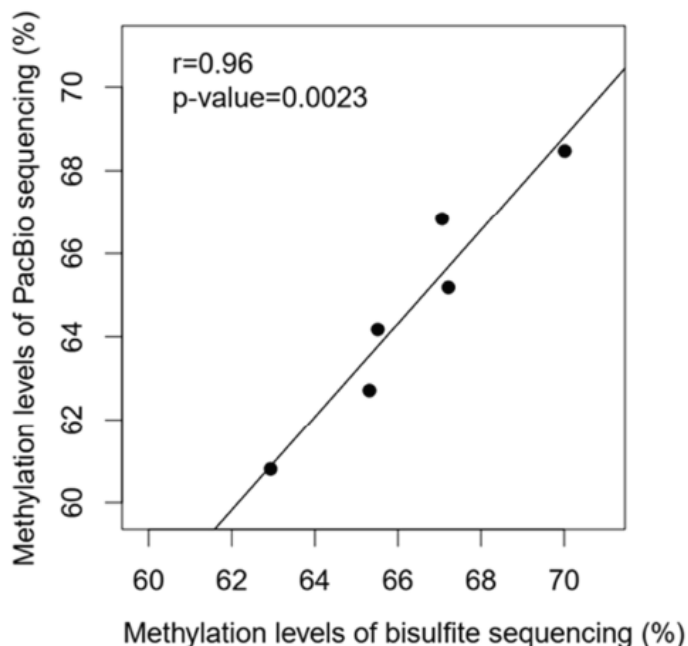
62. The ’794 patent demonstrates its method as recited in claim 1 can detect modifications at a level comparable to bisulfite sequencing. Specifically, the ’794 patent expressly compared 5mC modification detection using bisulfite sequencing with an embodiment of the method recited in claim 1 of the ’794 patent. That comparison demonstrates a good correlation (characterized by an r above 0.7) of methylation quantifications of various biological samples between the two methods. EX1001, FIGs. 51, 52B, 53, 63A, 65A-B, 66A-B, 68A-B, 69A-B, 71A-B, 72A-B, 86A, and 99.

63. The '794 patent demonstrates good correlation between the methylation levels measured in cell-free DNA (cfDNA) by the claimed method and the conventional bisulfite sequencing. EX1001 at 47:36-48:63.

64. The '794 patent demonstrates “a good correlation (r: 0.96, p-value=0.0023) between the overall methylation levels [in plasma DNA]deduced by the [claimed] approach and [conventional]bisulfite sequencing.” EX1001 at 48:11-26, EX2001, FIG. 51.

The y-axis is the methylation levels predicted according to the embodiments present in this patent application. The x-axis is the methylation levels deduced by bisulfite sequencing. A median of 314,675 CpG sites (range: 144,546-1,382,568) were analyzed for plasma DNA results generated from single molecule, real-time sequencing. The median proportion of CpG sites that were predicted to be methylated was 64.7% (range: 60.8-68.5%), which appeared to be comparable with results deduced from bisulfite sequencing. As shown in FIG. 51, there was a good correlation (r: 0.96, p-value=0.0023) between overall methylation levels deduced by the single molecule, real-time sequencing with the present methylation prediction approach and bisulfite sequencing.

EX1001 at 48:13-26,



EX2001, FIG. 51.

65. The '794 patent demonstrates good correlation ($r=0.97$, $P\text{-value} = 0.007$) between the genomic representations (GR) of Y chromosome in maternal plasma of pregnant women measured by the claimed approach and conventional bisulfite sequencing. EX1001 at 48:51-55, EX2001, FIG. 53.

66. The '794 patent recites that “[t]hese results suggested that single molecule, real-time sequencing would also allow the accurate quantification of DNA molecules originating from nonhematopoietic tissues such as placenta, whose contributed DNA generally represented a minority. In other words, this disclosure demonstrated the feasibility for simultaneously analyzing copy number aberrations

and methylation states for native molecules without any base conversions and amplifications prior to sequencing.” EX1001 at 48:55-63.

67. In summary as demonstrated in the '794 patent, the claimed modification detection method of claim 1 provides a novel approach for methylation detection using SMRT sequencing data of native nucleic acid molecules without base conversions or amplifications before sequencing.

68. The '794 patent compared regional methylation level quantification performed by the claimed approach and the conventional bisulfite sequencing.

69. The '794 patent showed good correlations ($r = 0.9$ or higher) between the claimed approach and the conventional bisulfite sequencing on quantifying overall methylation levels of DNA samples from placenta, buffy coat, a hepatocellular carcinoma (HCC) cell line, HCC tumor tissue and adjacent normal tissue (FIG. 63A, $r = 0.99$, P value < 0.0001 , EX1001 at 56:21-39). on methylation levels of a hepatocellular carcinoma (HCC) cell line (HepG2) measured at 1-Mb resolution (FIG. 65A $r = 0.99$, P value < 0.0001 , EX1001, 58:4-19 and at 100-kb resolution (FIG. 66A, $r = 0.98$, P value < 0.0001 , EX1001, 58:20-42); and on methylation levels of HCC tumor tissues measured at 1-Mb resolution (FIG. 68A, $r = 0.96$, P value < 0.0001 , EX1001, 58:66-59:14; and FIG. 71A, $r = 0.98$, P value < 0.0001 , EX1001, 59:54-60:2)) and at 100-kb resolution (FIG. 69A, $r = 0.9$, P value

< 0.0001, EX1001, 59:16-30; and FIG. 72A, $r = 0.94$, P value < 0.0001, EX1001, 60:3-17).

70. The '794 patent showed correlations between the two methods for methylation quantification of samples from buffy coat samples at 1-Mb resolution (FIG. 65B, $r = 0.87$, P value < 0.0001, EX1001, 58:18-19) and at 100-kb resolution (FIG. 66B, $r = 0.81$, P value < 0.0001, EX1001, 58:20-42); and adjacent normal liver tissue at 1-Mb resolution (FIG. 68B, $r = 0.83$, P value < 0.0001, EX1001, 59:9-15; and FIG. 71B, $r = 0.87$, P value < 0.0001, EX1001, 59:64-60:2) and at 100-kb resolution (FIG. 69B, $r = 0.75$, P value < 0.0001, EX1001, 59:16-30; and FIG. 72B, $r = 0.78$, P value < 0.0001, EX1001, 60:3-17).

71. The '794 patent compared the correlation of overall methylation levels quantified by bisulfite sequencing and the claimed method using Sequel II Sequencing Kit 2.0 (SMRT-seq). Ex1001 at 70:1-71:2. A good correlation between the two methods was shown in FIG. 86A ($r = 0.95$, P value < 0.0001). The differences between the two methods increased when the average of methylation levels measured increased. FIGs. 86B and 86C. The '794 patent states that the “error may be with bisulfite sequencing and not related to the methods with SMRT-seq” due to the biased sequence output and over-estimated global methylation introduced by bisulfite conversion. EX1001, 70:59-65. As the claimed method does not require bisulfite conversion before sequencing, the claimed method can be more

accurate in methylation detection than bisulfite sequencing by avoiding drastic DNA degradation caused by bisulfite conversion. EX1001, 70:66-71:2. I agree with these statements.

72. The '794 patent compared overall methylation levels determined by the claimed approach and bisulfite sequencing with the use of different subread depth cutoffs. EX1001, 79:17-46, FIG. 99. The correlation between these two measurements was well correlated ($r = 0.8$, $P \text{ value} < 0.0001$) when covered by subreads at least once, and the correlation improved when the subread depth cutoffs were increased to 10x ($r = 0.87$, $P \text{ value} < 0.0001$) and 30x ($r = 0.95$, $P \text{ value} < 0.0001$). EX1001, 79:27-46.

73. The '794 patent also includes experimental data supporting several 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), and methylation analysis in viral DNA (61:61-62:48). EX1001.

D. Prosecution History

74. I have reviewed the prosecution history of the '794 patent, including the Preliminary Amendment dated August 17, 2020 (EX1004 at 1090-95), the Office Action mailed December 30, 2020 (EX1004 at 795-812), the Applicant-Initiated

Interview Summary dated March 9, 2021 (EX1004 at 216-18), the Claim Amendments and Response to the Office Action mailed December 30, 2020 (EX1004 at 190-210).

75. The Applicant-Initiated Interview Summary dated March 9, 2021 under the heading 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 at 217. And under the heading 35 U.S.C. 103, it states: “See comments in 102 above.” *Id.*

76. I understand that the Examiner concluded Flusberg and the ’338 patent do “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.” *Id.* I agree with the Examiner’s conclusion that claim 1 is patentable over the Flusberg publications.

VII. TSE’S HOLISTIC KINETIC MODEL IS THE MODEL OF CLAIM 1.

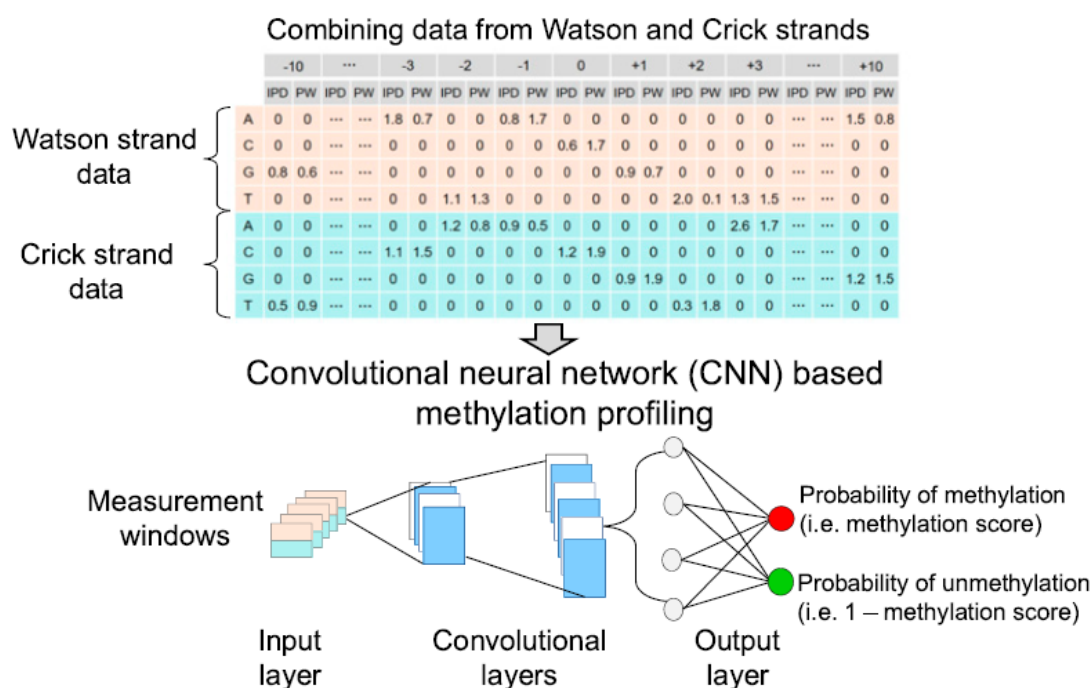
77. I have reviewed Tse et al., “Genome-wide Detection of Cytosine Methylation by Single Molecule Real-time Sequencing,” Proceedings of the National Academy of Sciences, 118(5) (2021) (“Tse”) (EX1030), which I understand is authored by the inventors of the ’794 patent and other collaborators.

In my opinion, Tse's holistic kinetic model is the model of claim 1 of the '794 patent and as shown in EX2054.

78. I have confirmed by searching on line that the Tse paper (Ex. 1030) is the only PNAS published in January 2021 with Dr. Dennis Lo as a co-author.

79. A portion of Fig. 1 from Tse is set forth below. A matrix, titled "Combining data from the Watson and Crick strands" shows a 21-nucleotide window, 10 nucleotides upstream and 10 downstream of a cytosine within a CpG site. The nucleotide identity of 21-base Watson strand was 5'-G[CCATGC]ATACGTT[GATGCA]A-3'. The nucleotides in the brackets were left out (denoted by "...") in the matrix. For a position of -3 corresponding to the base of adenine ("A"), the IPD (1.8) and PW (0.7) associated with "A" were filled in the corresponding cells between a column of "-3" and a row of "A." The other cells in the same columns were filled by "0." The remaining IPDs and PWs related to the 21 nucleotide strand were filled in that measurement window based on the same rule. The kinetic signals and nucleotide identity originating from the Crick strand (5'-T[TTGCAT]CAACGTA[TGCATG]G-3') were recorded similarly. The measurement windows for two CpG sites complementary to each other (i.e., the Watson strand and the Crick strand) were combined for downstream analysis. A number of combined measurement windows originating from methylated and unmethylated cytosines were used for training a CNN, so as to differentiate

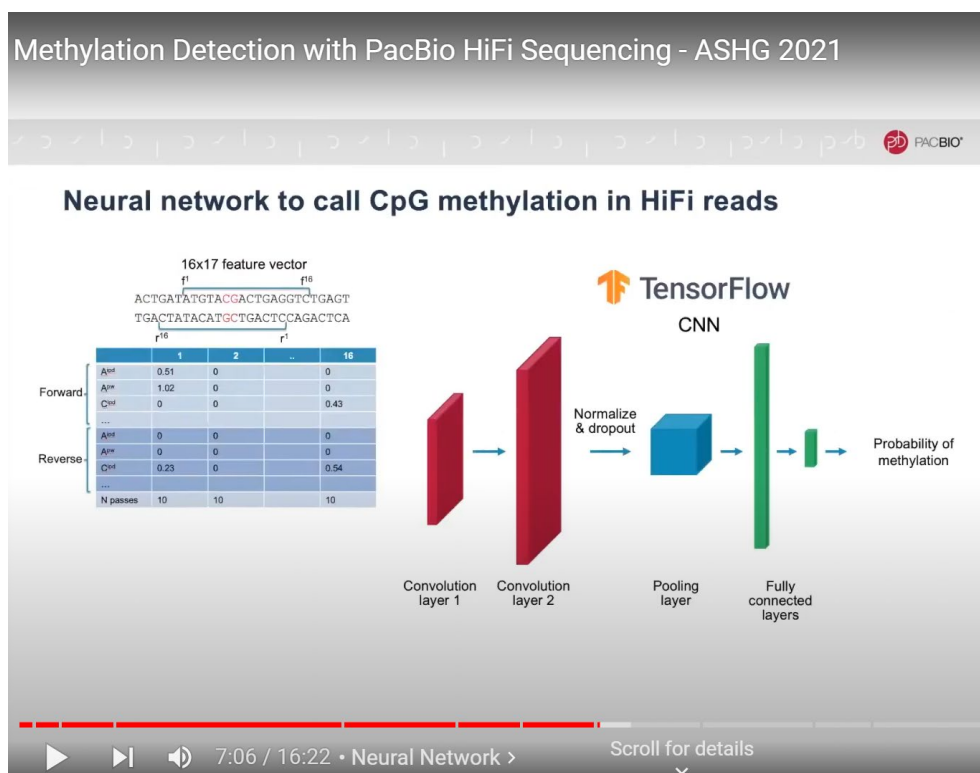
methylated and unmethylated cytosines in test samples. CNN involved input layer, convolutional layers, and output layer. The measurement windows were fed into the input layer, followed by the process of convolutional layers; then, the probability of methylation (range: 0 to 1) for a CpG was generated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



EX1030, Fig. 1 (partial).

80. I have watched a YouTube video (EX2024) of Dr. Aaron Wenger, PacBio’s Associate Director of Product Marketing, presenting “Methylation Detection with PacBio HiFi Sequencing.” Dr. Wenger described “models that integrate kinetic information over multiple bases and passes of a molecule” that “can provide high precision and recall for 5mC from PacBio HiFi reads.” *Id.*, 15:46–

15:56. Dr. Wenger's presentation showed a matrix as an input data structure for CNN modeling, trained by methylated and unmethylated samples, and accomplished 90% precision with 15-20 passes. *Id.* at 7:07-10:26, 14:38-end.



EX2024 screenshot at 7:06.

81. The above slide from Dr. Wenger's presentation bears a remarkable similarity to Fig. 1 of the Tse paper. The data matrix included the kinetic data from the forward and reverse strands which corresponded to the Watson and Crick strands of Tse, respectively. The matrix also displays the kinetic values (IPD and PW) for every nucleotide in the window (16 nucleotides) which is five nucleotides shorter than the Tse window. Visually, the difference between the two matrices is that Tse

uses two columns for each position, -10 IPD/PW, ..., -3 IPD/PW, -2 IPD/PW, etc.) to record the IPD and PW values.

-10	...	-3	-2	-1	0	+1	+2	+3	...	+10	
IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW

EX1030, Fig. 1 (partial).

82. Whereas Dr. Wenger uses two rows A^{IPD} , A^{PW} , etc.

A^{ipd}
A^{pw}
C^{ipd}
...
A^{ipd}
A^{pw}
C^{ipd}
...
N passes

EX2024 screenshot at 7:06 (partial).

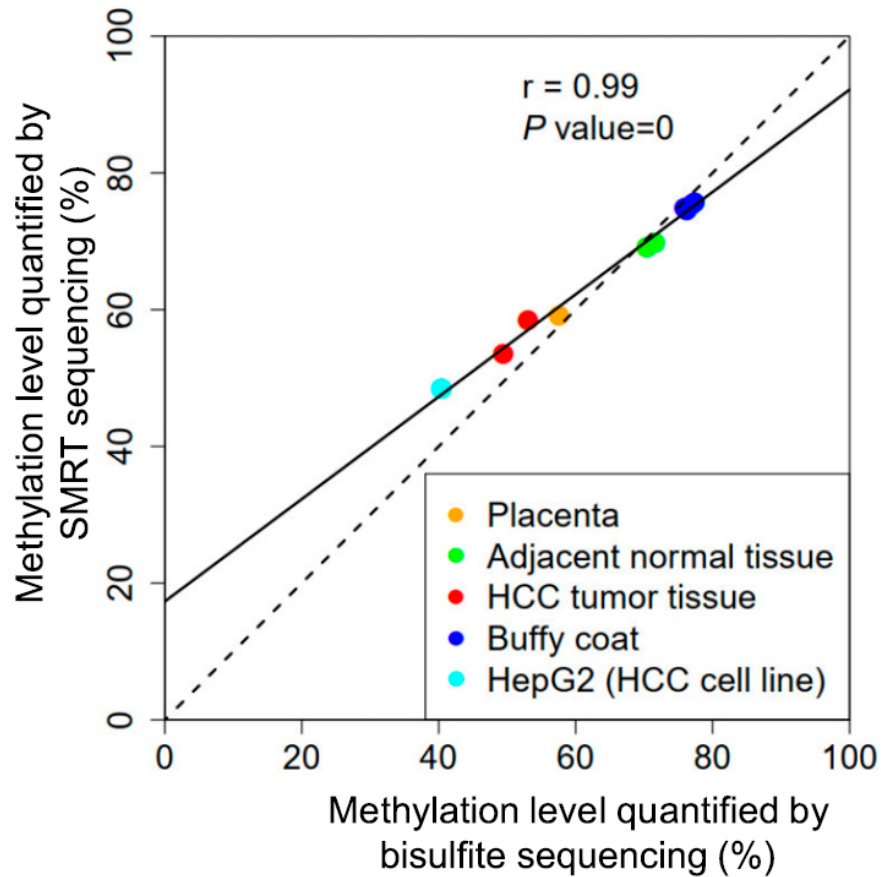
83. While the differences in organization of the matrix appear visual different to the reader, they are similar with respect to how a computer model processes the information for modification classification.

84. Dr. Wenger's matrix includes a row for the number of passes which refers to the subread depth. It is unclear from the video if these subreads were part of the input data structure of the model. However, Tse also analyzed the number of passes or subreads and found that "[t]he increase of subread depth would generally increase AUC values in differentiating methylated an unmethylated cytosines," which plateaued for the 21-nucleotide window at 30x. EX1030 at 5.

85. Dr. Wenger's slide identifies "TensorFlow CNN" which is a well known open source convolutional neural network. Keras is another well known open source open source convolutional neural network which was used in Tse. The two CNNs are comparable systems. Dr. Wenger's slide depicts the convolutional layers as red rectangles which correspond to the blue and white rectangles in Tse Fig. 1. EX1030, Fig. 1. Dr. Wenger's slide includes a pooling layer which is also disclosed in Tse. EX1030 at 10 ("A maximum pooling layer with a pool size of 2 was used."). Similar to Tse, Dr. Wenger's slide includes connecting layers depicted as green rectangles which in Tse are white circles shown as connected to the green and red output layers. EX1030 at 10 ("A flattened layer was further added, followed by a fully connected layer comprising 10 neurons with the use of the ReLU activation function. The output layer with one neuron was finally applied, with a sigmoid activation function to yield the probabilistic score for a CpG site of being methylated (i.e., methylation score)."). Lastly, Wenger's slide depicts a probability of methylation, just like Tse. EX1030 at 10, Fig. 1.

86. I reviewed Exhibit 2054 comparing Claim 1, Tse (EX1030), Dr. Wenger's presentation (EX2024), and a poster describing a later PacBio release of a product called primrose (EX2047). In my opinion, the DNA methylation detection tool disclosed in Dr. Wenger's presentation, and primrose have strong similarities to Tse's holistic kinetic model and have duplicative features.

87. In addition, Tse showed excellent correlation ($r = 0.99$, P value < 0.0001) of overall methylation determination in 11 tissue DNA samples between bisulfite sequencing and the holistic model. EX1030 at 6, Fig. 4.



EX1030, Fig. 4.

VIII. CLAIM 1 IS NOT OBVIOUS OVER FLUSBERG, THE '338 PATENT, AND THE PACBIO-BAM.

A. Flusberg does not Disclose or Suggest to a Person of Ordinary Skill in the Art the “Input Data Structure” of Claim 1.

88. Flusberg suggests a method of nucleotide modification detection that controls for nucleotide identity by keeping it constant. As such Flusberg does not disclose or suggest including nucleotide identity in an “input data structure” as claimed.

89. 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 at 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.

90. 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 at 462. In this passage, the term “sequence context” does not

refer to input data, but rather the known window around a base (adenine) modification.

91. As explained in EX2002, ¶63, Flusberg’s kinetic data analysis focused on IPD ratios that correlated with the location of known methylated bases. EX1006 at 462, Fig. 2.

92. 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 at 462-63. Again, the use of the term “sequence context” in Flusberg is not as input data, but rather an observation regarding a pattern of IPD ratios associated with the nucleotide positions surrounding a known methylation site, *i.e.*, a window.

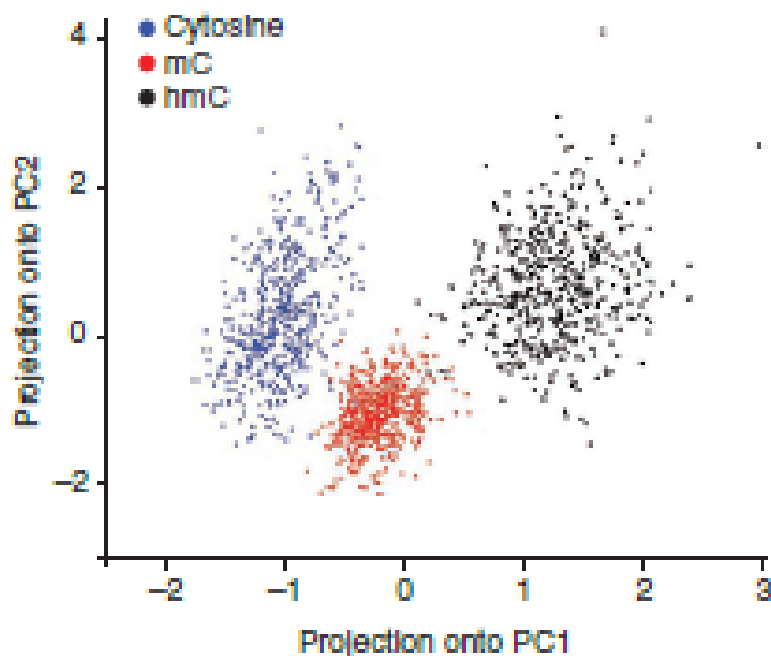
93. Flusberg also measured the pulse width ratios and found that for the same modification, the pulse width ratios were significantly lower than the IPD ratios, and the IPD ratio also exhibits substantially higher signal-to-noise ratio than the pulse width. See EX2002 ¶¶66-67. IPD ratios had a substantially higher signal-

to-noise ratio (2-3X) defined by the ratio of the dynamic range around a modified base compared to background positions.

94. 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 EX2002 ¶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 at 130.

95. 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 at 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.

96. Fig. 3 in Flusberg 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. EX1006 at 463, Fig. 3.

97. 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 method of Flusberg resulted in an AUC of 0.55 which is essentially random guessing. EX2001, FIG. 29.

98. The next section of Flusberg titled Results: Methylation detection by circular consensus sequencing (EX1006 at 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.

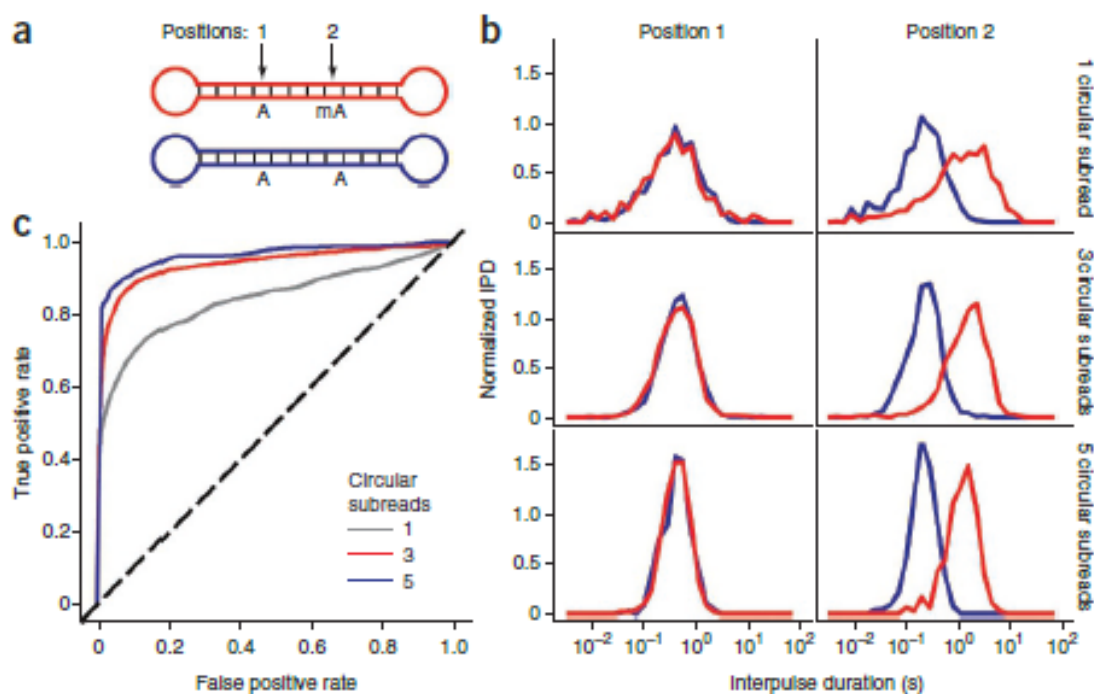


Figure 4 | IPD distributions for adenine and mA in synthetic DNA templates. (a) Schematic of the 199-base DNA templates used in this experiment. (b) IPD distributions at the indicated positions for both templates, colored as in a. The histograms depict the distributions of mean IPD (averaged over the 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.

99. 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.

100. Flusberg then used SMRT sequencing and compared the IPD kinetics between the unmethylated and methylated sequences (6mA in GATC motifs) as explained in EX2002 ¶73; EX1006 at 464-65.

101. As explained in EX2002, ¶¶73-74, Flusberg observed that IPD ratio in methylated DNA samples over the control unmethylated DNA sample provided better separation between methylated motifs (GATC, black squares) and unmethylated motifs (Non-GATC, open circles) than average IPD values. EX1006, Figs. 5b-5d. The average IPD values of methylated motifs were higher than a majority of non-methylated motifs, there remained non-methylated motifs that had comparable or even higher IPD than methylated motifs. EX1006, Figs. 5b and 5c.

102. 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 at 2912.

103. Accordingly, 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.

104. 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	<i>P</i> 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^{-19}
2261	GGATCAGATCCAATTA	5.56	1.43×10^{-22}
2499	CAGATGGATCAATCAA	2.43	4.44×10^{-7}
2583	ATTTTGATCTAGTTT	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.

105. 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 at 467. Here, Flusberg uses “local context” to refer to a four-base motif that was under investigation, not using nucleotide identity as input data.

106. 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, ¶107. There is no teaching or suggestion to a person of ordinary skill in the art in Flusberg to include nucleotide identity as a property in an "input data structure" as claimed.

B. The '338 Patent does not Disclose or Suggest to a Person of Ordinary Skill in the Art the "Input Data Structure" of Claim 1.

107. Like Flusberg, the '338 Patent discloses a method of nucleotide modification detection that controls for nucleotide identity by keeping it constant. As such the '338 patent does not disclose or suggest including nucleotide identity in an "input data structure" as claimed.

108. 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.

109. The disclosure of the '338 patent covers a variety of topics as summarized in EX2002, ¶¶79-101. Like Flusberg, the methyladenosine experiments of the '338 patent were 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. Like Flusberg, the '338 patent's reference to sequence context was observational - that the IPD around a modified base differed several bases upstream and downstream of the modified base, *i.e.*, a window.

110. In sum, all the experiments in the Examples sections of the '338 patent 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.

111. The '338 Patent does not include or suggest the inclusion of nucleotide identity in an input data structure as claimed. 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 was a constant in the experiments that observed polymerase kinetics over a sequence context, not a variable. In other words, the '338 Patent discloses a method of detection that controls for nucleotide identity.

112. 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, ¶¶101-20.

113. 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.” EX1002, ¶103 (also cited in ¶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.

114. 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 suggest to a person of ordinary skill in the art to use nucleotide identity as property in an input data structure as claimed. To the contrary, a POSA would not modify a method that controls for nucleotide identity by keeping it constant and thus eliminating it a variable, and

instead introduce an addition property into the input data structure, further complicating the analysis and complexity of the system.

115. Dr. Dessimoz addresses “the identity of each nucleotide” in paragraph 111 of his declaration. EX1002, ¶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 ’338 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.

116. 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 including nucleotide identity as a property in an input data structure as recited in claim 1.

117. Moreover, the above passage refers to the relationship between pausing and a modification.

118. Dr. Dessimoz next cites the '338 patent (EX1007), 64:53-55 which states: "It [i.e., 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." EX1007, ¶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.

119. Moreover, this passage refers to using base calls for detecting the likelihood of pausing. EX1007, 64:28-55. The '338 patent further discloses that this likelihood of pausing can be used to help detect a modification. EX1007, 65:13-16. This disclosure regarding methods of detecting likelihood of pausing does not disclose using nucleotide identity as a property in an input data structure as recited in claims 1. Just like the '338 patent's method of controlling for nucleotide identity, this disclosure relates to a different way to account for nucleotide in base modification detection, namely, first find likelihood of pausing then use that result to inform modification detection.

120. 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, ¶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.

121. As the above passage explains, the two templates being compared were identical which means the nucleotide identity was a constant, not a variable. Consequently, the passage confirms that the '338 patent controlled for sequence context. A POSA would not understand that passage as disclosing or suggesting a different way to account for sequence context and a POSA would not be motivated to modify such a system to a more complex and resource intensive system by

introducing nucleotide identity as a property in an input data structure as recited in claim 1.

122. 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, ¶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 SMRT™ 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, SMRT™ 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.

123. Again, because the DNA templates had an identical sequence, nucleotide identity was a constant not a variable.

124. 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, ¶111. As explained above, the disclosure at 64:29-65:20 is directed to a model for differentiating between pauses and sequencing states, not detecting 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.

125. 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 include nucleotide identity in an input data structure as claimed. The input for nucleic acid modification detection may be "the conditional probability of a signal arising from a pause or sequencing state given the observed pulses" instead of the "observed pulse data (both basecalls and other pulse features)." Or alternatively, as Dr. Dessimoz stated in his deposition, the point is to "figure out if

you've got pauses[, a]nd if you do, then you can send it somewhere else, to another model, to figure out if it's a modification” Dessimoz deposition transcript (EX2037), at 121:4-18.

126. Lastly, Dr. Dessimoz cites the '338 patent (EX1007) at 7:4-17 as a “pause indicative of modification.” EX1002, ¶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.

127. In sum, the disclosure of the '338 patent does not teach or suggest to a POSA to include nucleotide identity as a property in “an input data structure” as claimed. To the contrary, the '338 Patent, just like Flusberg, discloses a method that controls for nucleotide identity by keeping it constant.

C. The PacBio-BAM does not Teach or Suggest to a Person of Ordinary Skill in the Art the “Input Data Structure” of Claim 1.

128. 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), 2078-79 (2009) (EX2004).

129. The PacBio BAM “format is a binary, compressed, record-oriented container format for raw or aligned sequence reads.” EX1011 at 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-ready subreads from movie,” etc.). EX1011 at 2. The PacBio BAM file itself cannot be used as the claimed “input data structure” because the data in the BAM 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.

130. 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.

131. 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 at 8. Nucleotide identity is not a kinetic pulse feature. And, the PacBio BAM does not disclose or suggest any particular method for base modification detection. Instead, the PacBio-BAM indicates that its format saves disk space without affecting the kinetic data (IPD and PW) for use cases like "base modification detection." *Id.* Based on the disclosure in the PacBio-BAM, a POSA would not be motivated to modify the methods of the Flusberg publications, that control for nucleotide identity by keeping it constant, and instead include nucleotide identity as a property in an input data structure as claimed.

D. A Person of Ordinary Skill in the Art Would not be Motivated to Combine Flusberg, the '338 Patent, and the PacBio-BAM to Arrive at the Invention of Claim 1.

132. Flusberg and the '338 patent do not teach or suggest the use of nucleotide identity as a property in an "input data structure" as claimed as I have explained above. The PacBio-BAM file is the stored data generated in the normal course of SMRT sequencing and does not teach or suggest any particular method of modification detection. A POSA would not be motivated to modify the methods of the Flusberg publications that control for nucleotide identity, and instead include nucleotide identity as a property in the claimed input data structure.

133. I have reviewed Dr. Dessimoz's declaration regarding the combination of Flusberg, the '338 patent, and the PacBio BAM file. EX1002, ¶¶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 publications 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, the '338 patent, EX1007, Clark (2011), EX1008, Loomis (2013), EX1009, Schadt (2013), EX1010, and PacBio's White Paper EX1016. But, none of

those papers discloses including the identity of a nucleotide in an input data structure for base modification as claimed. To the contrary each of those references either follows the proposed methods of the Flusberg publications that control for nucleotide identity by making it constant, or does not include any method of detection a nucleotide modification.

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 at 130.

137. Similar to Flusberg, Clark (EX1008) controlled for nucleotide identity by making it constant and did not include nucleotide identity in an "input data structure" as claimed. Specifically, Clark 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 (~20-40X increase), and O4-methylthymine (~10-30X increase) modifications at the site of the modifications. EX1008 at 2 and 5, Figures 1-2. But, none of the corresponding pulse width ratios showed a strong correlation.

138. Dr. Dessimoz cites Loomis (EX1009) as “corroborating the obviousness of considering data for nucleotide identity,” but Loomis does not disclose or suggest how to use nucleotide identity to detect nucleotide modifications. Nor does Loomis disclose or suggest how to use IPD or PW to detect nucleotide modifications. Loomis cites to Flusberg for “the previously described context effect on polymerase kinetics (Flusberg et al. 2010).” EX1009 at 126. Loomis then suggests that its “observations thus establish a clear foundation for the detection of epigenetic modifications within the CGG-repeat region, which has not been possible at the nucleotide level for full mutation alleles that are epigenetically silenced in FXS.” EX1009 at 126. Here, Loomis makes the same observation as the Flusberg publications and suggests that modification detection within the “CCG region” may be possible. Loomis does not disclose a method of detecting a modification or how nucleotide identity, PW or IPD should be considered in such a method. And, Loomis’ analytical methods do not use PW.

139. The papers that followed, Schadt (2013), EX1010, and PacBio's White Paper EX1016, also all controlled for nucleotide identity by making it constant and did not include nucleotide identity in an "input data structure" as recited in claim 1. Those references also 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) at 130.

140. Schadt "developed a model that considers IPD information at sites flanking the site of interest and allows for neighboring sites to interact. . . ." EX1010 at 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 at 130.

141. 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 – Clark (2011), EX1008, and Schadt (2013), EX1010, using an IPD model reinforces that teaching, as does PacBio’s White Paper (EX1016) and kineticsTools as described in the PacBio kineticsTools manual (EX1021). Because of those teachings, by 2019 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.

142. PacBio’s White Paper (EX1016) also controlled for nucleotide identity by making it constant and did not include nucleotide identity in an “input data structure” as recited in claim 1. EX1016 at 3. PacBio’s White Paper (EX1016) discloses using IPD, but does not disclose using PW. PacBio’s White Paper describes two techniques for detecting DNA modifications both using IPD as the input: A comparison of IPD ratio between a DNA sample that is divided into native templates and whole-genome amplified control templates or a comparison is made between a native sample and an “in silico control [that]uses a predicted value of the mean IPD at each reference position that is based on the local sequence context.” EX1016 at 3-4. Figure 4 of PacBio’s White Paper illustrates the IPD kinetic signature detected by this method with IPD ratio plotted on the Y-axis v. template

position on the X- axis. Notably, PacBio's White Paper omitted pulse width as an input in either technique.

143. Similarly, PacBio's kineticsTools as described in the PacBio kineticsTools manual (EX1021) also controlled for nucleotide identity by making it constant and did not include nucleotide identity in an "input data structure" as recited in claim 1. EX1021 at 1. Like PacBio's White Paper, the PacBio kineticsTools manual discloses using IPD, but does not disclose using PW. EX1021 at 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." As in PacBio's White Paper, "[t]he expected IPD value for unmodified DNA can come from either an *in silico control* or an *amplified control*." *Id.* Notably, the PacBio kineticsTools manual does not disclose kineticsTools using pulse width as an input to detect DNA modifications.

144. Dr. Dessimoz opines that kinetic parameters and sequence context were used in other contexts, specifically Oxford Nanopore sequencing.

145. First, I note that claim 1 is expressly limited to "receiving data acquired by measuring pulses in an optical signal corresponding to nucleotides sequenced in a sample nucleic acid molecule." EX1001, 115:40-43. As such, claim 1 expressly excludes Oxford Nanopore sequencing data, which does not measure pulses in an optical signal, and in fact does not include an optical signal at all. Instead, Oxford

Nanopore sequence measure electrical signals. EX2045 at 2 (“measures the electric current that flows through the nanopore.”).

146. Moreover, as explained in the '794 patent while Oxford Nanopore sequencing allows for long reads 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 Biosciences 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.

147. 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 at 3-4. The approaches are not comparable.

148. Moreover, Oxford Nanopore sequencing does not involve polymerase kinetics at all. Dr. Dessimoz's opinion relies on the relationship between sequence context and polymerase kinetics to assert that a POSA would combine nucleotide identity with the polymerase kinetic properties IPD and PW. But, there is no similar or equivalent relationship applicable to Oxford Nanopore sequencing. That nucleotide identity was used in Oxford Nanopore sequencing does not teach or motivate a POSA to use nucleotide identity in PacBio SMRT sequencing or combine nucleotide identity with polymerase kinetic parameters like PW and IPD.

149. In summary, a POSA would understand that including nucleotide identity as a property in an input data structure for base modification detection or to combine nucleotide identity with PW and IPD was not well-characterized or known in 2019. To the contrary, the well-known use of nucleotide identity in detecting base modification with either PW and/or IPD at the time was to control for nucleotide identity by making it constant. That a POSA might have understood that sequence context or even nucleotide identity should be considered in nucleotide modification detection does not suggest to a POSA which properties related to sequence context

should be considered or how to account for any one of those properties including how to use nucleotide identity.

150. Moreover, in 2019 a POSA would not understand or have reason to believe that the method of the Flusberg publications, and the several references that follow that same approach as Flusberg as explained above, that control for nucleotide identity by making it constant would be interchangeable with the claimed method that includes nucleotide identity as a property in an input data structure as recited in claim 1.

151. Therefore, by 2019, a POSA would not have combined nucleotide identity, PW and IPD as properties in an “input data structure” as claimed in view of the series of references described above that lead a POSA down a different path, to focus on IPD and not include PW in the method of detection.

IX. OBJECTIVE INDICIA OF NON OBVIOUSNESS

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

152. I understand that the objective indicia of long-felt unsolved need is 1) a persistent need recognized by POSAs, (2) that was not satisfied by another before the invention by the inventor, and (3) the invention satisfies the long-felt need.

153. While bisulfite sequencing has been considered the gold standard technology for DNA methylation analysis, bisulfite sequencing has significant

drawbacks as explained above in the '794 patent. Both Flusberg and the '338 patent recognized similar drawbacks to bisulfite sequencing:

Despite these advances, enabled by bisulfite treatment-based sequencing, there are several drawbacks to the technique. For example, the sample preparation steps associated with bisulfite treatment can be costly and time-consuming, and the harsh reaction conditions necessary for complete conversion can degrade DNA. Additionally, the reduction of complexity in converted genomes constrains primer design for subsequent PCR amplification and complicates alignment to a reference genome. Finally, discrimination between cytosine, mC and hmC cannot be accomplished with bisulfite treatment-based sequencing.

EX1006 at 461, see also the '338 patent, EX1007, 2: 39-49, 68:62-69:2.

The '338 patent explains additional drawbacks of bisulfite sequencing with respect to *de novo* methylation profiling.

Furthermore, because bisulfite sequencing relies on either microarray or 2nd-generation DNA sequencing technologies for its readout of methylation status, it also suffers from the same limitations as do these methodologies. For array-based procedures, the reduction in sequence complexity caused by bisulfite conversion makes it difficult to design enough unique probes for genome-wide profiling. Most 2nd-generation DNA sequencing technologies employ short reads and thus have difficulties aligning to highly repetitive genomic regions. This is especially problematic, since many CpG islands reside in such regions. Given these limitations, bisulfite sequencing is also not well suited for *de novo* methylation profiling.

EX1007, 2:50-63 (citations omitted).

154. Flusberg identifies the need by stating that: “[w]e describe the direct detection of DNA methylation, without bisulfite conversion, through single-molecule, real-time (SMRT) sequencing.” EX1006 at 461. As explained above Flusberg was a proof-of-concept paper. Flusberg did not solve the need for “the direct detection of DNA methylation, without bisulfite conversion, through single-molecule, real-time (SMRT) sequencing” because, as explained by Flusberg’s co-authors three years later in Schadt et al., 5-mC detection required bisulfite conversion and SMRT sequencing was not a surrogate for chemical modification of nucleic acids.

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

EX1010 at 129 (citation omitted).

155. Schadt et al. explained that “no statistical framework has been proposed to identify kinetic variation events from SMRT sequencing data that can serve as a surrogate for chemical modifications of DNA or RNA sequences.” EX1010 at 130. And, at the time the most common chemical modification of DNA for methylation detection was bisulfite sequencing.

156. That need persisted until the '794 patent was filed which used the method of claim 1 to directly detect DNA methylation, without bisulfite conversion, through SMRT sequencing and produced comparable results to bisulfite sequencing. (See above at ¶¶59-72).

157. Between the Flusberg publications and '794 patent, PacBio developed a software program for detecting DNA modifications without bisulfite conversion that used IPD, not PW, and controlled for nucleotide identity by holding it constant such that neither nucleotide identity nor PW was included in an input data structure as recited in claim 1. That program called “kineticsTools” did not solve the long-felt need as reflected by the publications of multiple authors, including PacBio authors, acknowledging the problems of direct detection of DNA methylation, without bisulfite conversion, through SMRT sequencing. Beaulaurier (EX2038) at 159 (“[D]etection of 5mC and 5hmC requires either high sequencing depth or additional steps to convert them to 5-formylcytosine and 5caC, both of which have higher signal-to-noise ratios.”); Gouil (EX2040) at 643 (“[T]he subtle effect of 5mC and 5hmC increase the [strand coverage] requirements to 250X, unless they are enriched or modified to produce a larger kinetic effect by glycosylation of TET-

conversion to 5-carboxylcytosine.² PacBio sequencing therefore only achieves single-molecule resolution for certain marks and on relatively short fragments(≤ 2 kb) that can be read a large number of times by the polymerase.”); Gigante (EX2051) at 2 (“We note that, for the dominant eukaryotic genome base modification at 5mC, the PacBio technology requires very high coverage making it impractical for use in the analysis of mammalian genomes. PacBio SMRT sequencing can be combined with bisulfite treatment (SMRT-BS) to facilitate 5mc detection, but this approach is currently only available for targeted sequencing and bisulfite treatment introduces the same drawbacks noted above in addition to fragmentation the DNA.”). In sum, a POSA would understand that, as recognized in the art, prior to the invention of claim 1 in the ’794 patent, the long-felt need for direct detection of DNA methylation, without bisulfite conversion, through SMRT sequencing had not been satisfied.

158. As explained above, the ’794 patent tested various features as properties to be used in the input data structure and concluded that the combination of IPD, PW, and nucleotide identify and relative position to a target position in a window

² While TET-conversion is technically not bisulfite conversion, it still precludes direct detection of DNA methylation through SMRT sequencing because it requires enzymatic conversion of the DNA template prior to sequencing.

was the most predictive feature set of DNA methylation detection. Using these features, the identity of the nucleotide, the position of the nucleotide relative to a target position in a window, the PW corresponding to the nucleotide and the IPD, the '794 patent inventors, the Lo inventors, trained their model according to the method of claim 1. EX2002 at ¶¶44-48, EX1001 at 30:60-31:63. Using the trained model with the input data structure of claim 1, as explained above, the Lo inventors conducted a series of experiments that compared their method to the “gold standard” – bisulfite sequencing and showed good correlation ($r = 0.90$ or higher) for most of the tested biological samples. (See ¶¶59-72 above). Dr. Lo's team solved the long-felt need for direct detection of DNA methylation, without bisulfite conversion, through SMRT sequencing.

159. Additionally, the method of claim 1 significantly lowered the subread depth required for 5mC detection using SMRT without bisulfite conversion. PacBio's White Paper recommended the minimum sequencing fold coverage per strand as 250x for 5mC (EX1016 at 4) and the art found the 250x minimum sequencing fold coverage per strand limited the tool to relatively short fragments. Gouil et al., EX2040 at 643. The '794 patent specification disclosed comparisons of overall methylation levels determined by the method of claim 1 with levels determined by bisulfite sequencing with the use of different subread depth cutoffs. EX1001, 79:17-46, FIG. 99. The correlation between these two measurements was

well correlated ($r = 0.8$, $P \text{ value} < 0.0001$) when covered by subreads at least once, and the correlation improved when the subread depth cutoffs were increased to 10x ($r = 0.87$) and 30x ($r = 0.95$). *Id.*

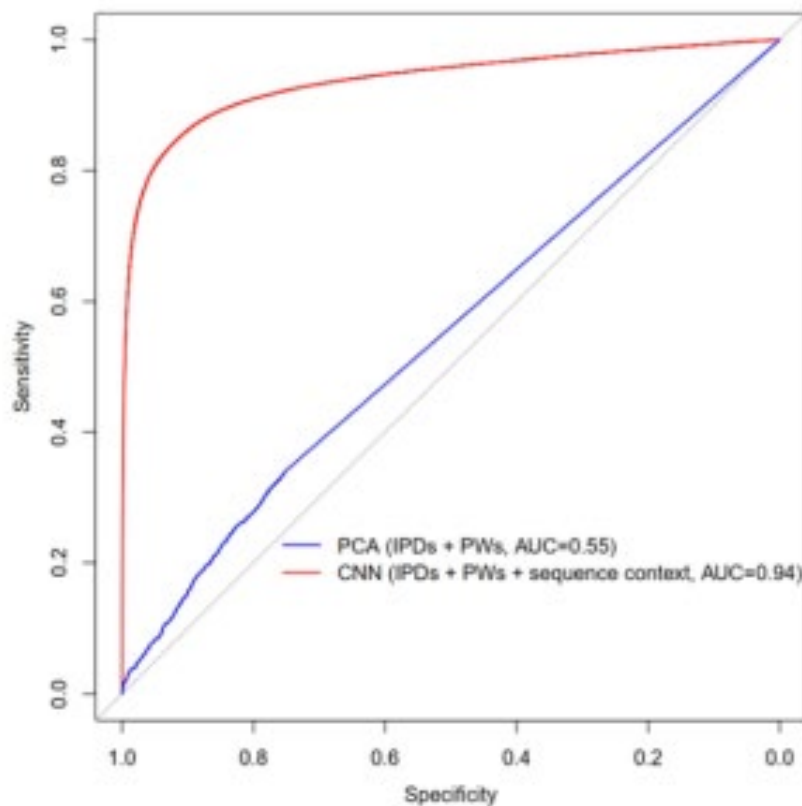
B. Unexpected Results

160. I understand that unexpected results must be shown to be unexpected compared to the closest prior art. Here the closest prior art is Flusberg publications, which controlled for nucleotide identity.

161. At the time the '794 patent was invented, a POSA could not predict the results of using the combination of features of input data structure of claim 1 of the '794 patent in light of the Flusberg publications, whether the results could be the same, better or worse. In order to know how well the model would perform with the features used in the input data structure of claim 1, a POSA would need to run experiments similar to those disclosed in the '794 patent specification that showed the properties of the input data structure of claim 1 were most predictive (*e.g.*, EX2001, FIGs. 18A & B).

162. Thus, a POSA would not have expected that modifying the methods of the Flusberg publications to arrive at the method of claim 1 would have improved the Flusberg methods to be comparable to bisulfite sequencing. In other words, it is surprising and unexpected that the method of claim 1 significantly improves the Flusberg publications' method such that the method of claim 1 is comparable to

bisulfite sequencing methods as demonstrated by the comparative data in the '794 patent.



EX2001, FIG. 29.

163. FIG. 29 shows that the AUC of the method based on the principal component analysis for which two principal components were used in Flusberg's study involving IPDs and PWs. As shown above, the Flusberg method was much less accurate (AUC: 0.55) than the method of claim 1, using a convolutional neural network involving IPD and PW as well as sequence context (AUC: 0.94). EX1001, 35:38-44.

164. The surprising and unexpected result of the method of claim 1 is its very high prediction (0.94) compared to the Flusberg method (0.55) which is essentially random guessing.

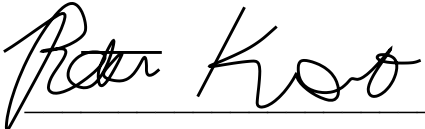
165. Additionally, as explained above, the method of claim 1 provides comparable results to bisulfite sequencing which is another unexpected result over the Flusberg publications. And, while Flusberg predicted that someday in future *de novo* methylation profiling may become possible with SMRT sequencing (EX1006 at 465), with the input data structure of claim 1, the Lo inventors were able to successfully use SMRT sequencing for *de novo* methylation profiling of a variety of cells which is another unexpected result in light of the Flusberg publications.

166. 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: July 26, 2024

By: _____

Peter K. Koo, Ph.D.