

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.

Case No. IPR2024-00028
Patent No. 11,091,794

DECLARATION OF DR. CHRISTOPHE DESSIMOZ

TABLE OF CONTENTS

I.	QUALIFICATIONS	1
II.	SCOPE OF WORK.....	4
III.	LEGAL STANDARDS	5
IV.	OVERVIEW	8
V.	BACKGROUND IN THE ART	12
	A. Third-Generation Sequencing	12
	B. SMRT Platform and Metrics Predictive of Modified Bases	13
	C. Predictive Models.....	21
	D. Deep Learning and Computing Power	26
VI.	PRIOR ART.....	29
	A. The Flusberg Publications (Flusberg and Flusberg-338).....	29
	B. PacBio-BAM.....	30
	C. Clement.....	33
	D. Template Guide	33
	E. Tsai	34
VII.	THE '794 PATENT	34
VIII.	PACBIO-BAM AND THE TAKE2 INFRINGEMENT CONTENTIONS	39
IX.	OVERVIEW OF THE LEVEL OF SKILL.....	40
X.	CLAIM CONSTRUCTION	41

XI.	GROUND 1 – CLAIMS 1-2, 7, 10-12, 14-15, AND 17-19 ARE OBVIOUS AS ILLUSTRATED BY THE FLUSBERG PUBLICATIONS AND PACBIO-BAM	41
A.	Claim 1	43
1.	Preamble.....	50
2.	Element 1.1	52
3.	Element 1.2	58
4.	Element 1.3	72
5.	Element 1.4	78
6.	Element 1.5	81
7.	Element 1.6	84
8.	Element 1.7	88
B.	Claim 2	90
C.	Claim 7	94
D.	Claim 10	98
E.	Claim 11	99
F.	Claim 12	101
G.	Claims 14 and 17.....	103
H.	Claim 15	106
I.	Claim 18	107
J.	Claim 19	108
XII.	GROUND 2 – CLAIMS 3-6 AND 9 ARE OBVIOUS AS ILLUSTRATED BY THE FLUSBERG PUBLICATIONS, PACBIO-BAM, AND CLEMENT.....	110
A.	Claims 3 and 4.....	112

B.	Claim 5	115
C.	Claim 6	116
A.	Claim 9	117
XIII.	GROUND 3 – CLAIMS 8 AND 15-16 ARE OBVIOUS AS ILLUSTRATED BY THE FLUSBERG PUBLICATIONS, PACBIO- BAM, AND THE TEMPLATE GUIDE	122
A.	Claim 8	123
B.	Claims 15 and 16.....	125
XIV.	GROUND 4 – CLAIMS 13 AND 15-16 ARE OBVIOUS AS ILLUSTRATED BY THE FLUSBERG PUBLICATIONS, PACBIO- BAM, AND TSAI.....	127
A.	Claim 13	128
B.	Claims 15 and 16.....	131
XV.	CONCLUDING STATEMENTS.....	132
XVI.	APPENDIX – LISTING OF CHALLENGED CLAIMS	134
XVII.	EXHIBIT LIST.....	138

I, Christophe Dessimoz, declare as follows:

I. QUALIFICATIONS

1. I have been retained by Counsel on behalf of Pacific Biosciences, Inc. to provide opinions in this matter related to my experience and expertise. My opinions are based on my skills, knowledge, training, education, and experience in matters of this nature, and my examination of the materials utilized in preparing this report.

2. I am a Professor of Bioinformatics at the University of Lausanne in Lausanne, Switzerland, and I am also the Executive Director of the Swiss Institute of Bioinformatics. My primary research interests include genomics, computer science, bioinformatics, and evolutionary biology.

3. In 2009, I received my Ph.D. from the ETH Zurich Department of Computer Science. I also earned a Diploma degree, equivalent to a Bachelor and Master of Science degree, from the ETH Zurich Department of Biology in 2003. I joined the University of Lausanne as SNSF Professor in 2015, and became an Associate Professor in the Department of Computational Biology in 2021. I was appointed the Executive Director of the Swiss Institute of Bioinformatics in 2022. Previously, I have worked at University College London, where I held the positions of Lecturer from 2013 to 2015, Reader from 2015 to 2018, Associate Professor from 2018-2020, and Professor from 2020-2022. I have also previously

worked as teaching staff at the University of Cambridge and as a research fellow at the EMBL-European Bioinformatics Institute. I also conducted graduate and postdoctoral research at the Department of Computer Science of ETH Zurich.

4. I have conducted extensive research in the field of bioinformatics. Some example research projects that I have conducted include devising computational methods to reconstruct evolutionary relationships among genes across different species, which is essential to be able to transfer biological knowledge from model species such as guinea pigs to human; assessing the accuracy of algorithms that operate on DNA sequences, such as sequence alignment methods; or developing algorithms to identify among a large number of DNA sequences those that represent fragments of the same gene.

5. I have authored or co-authored over 100 articles in peer reviewed journals or conference proceedings. Examples of such articles include evaluating the computational methods used to infer nucleotide sequences from raw signals in DNA sequencing (work published in the journal *Briefings in Bioinformatics* in 2011), storing arbitrary data in DNA (work published in *Nature* in 2013), devising computer methods to detect large alterations in DNA (work published in *Nature Communications* in 2017), or developing methods to identify the corresponding genes across different species (work published in *Nucleic Acids Research* in 2015).

6. I have received extensive research funding via grants from an array of sources, including the Swiss National Science Foundation, the European Science Foundation, the Swiss Institute of Bioinformatics, the University of Lausanne, and various private sources. My funding includes six active grants as well as 29 completed grants, totaling 14 million Swiss francs, or about \$15.5 million, with roughly half of that funding being grants for which I was the primary principal investigator.

7. I have taught various undergraduate and postgraduate courses at the University of Cambridge and University College London in the United Kingdom, and at ETH Zurich, the University of Basel, and the University of Lausanne in Switzerland. I currently teach a postgraduate course entitled “Bioinformatic algorithms,” where students learn techniques to analyze DNA sequences as well as machine learning approaches for classification, including neural networks. I have advised and chaired the dissertation of numerous doctoral and diploma students, including as the main supervisor of six Ph.D. students, a co-advisor of ten other Ph.D. students, and over 50 students at the bachelor’s or master’s level.

8. I have served as an editor and/or committee member of numerous professional organizations. For example, I have served on the managing board of *Peer Community in Mathematical and Computational Biology*, as a member of the Swiss National Science Foundation Council, as a member of the Board of

Directors of the International Society of Computational Biology, and as President of the Bioinformatics Intersection, Life Science Switzerland. I have also served as a grant reviewer for multiple institutions, including Institut Pasteur, the European Research Council, and the US-Israel Binational Science Foundation, as well as reviewing grants under national research councils and foundations of various nations, such as Switzerland, Canada, the UK, Belgium, and France.

9. A copy of my *Curriculum Vitae*, attached as EX1003, contains further details on my education, experience, publications, and other qualifications to render an expert opinion in this matter.

II. SCOPE OF WORK

10. I understand that a petition is being filed with the United States Patent and Trademark Office for Inter Partes Review of U.S. Patent No. 11,091,794 to Lo, et al. (“the ’794 patent,” attached as EX1001), entitled “Determination of Base Modifications of Nucleic Acids.” I have been retained by Pacific Biosciences of California, Inc. (“PacBio”) to offer an expert opinion on the validity of claims 1-19 of the ’794 patent assigned to The Chinese University of Hong Kong (“CUHK”). I receive \$475 per hour for my services. No part of my compensation is dependent on my opinions or on the outcome of this proceeding.

11. I have been specifically asked to provide my opinion on the unpatentability, of claims 1-19 of the ’794 patent. In connection with this analysis,

I have reviewed the '794 patent. I have also reviewed and considered various other documents in arriving at my opinions and may cite to them in this declaration.

III. LEGAL STANDARDS

12. I understand that a claim is not patentable under 35 U.S.C. §102 as anticipated if the claim as a whole is described, either expressly or inherently, in a single prior art reference. I understand that this also requires that the prior art reference disclose that each claim element is arranged as recited in the claims.

13. I understand that differences between the prior art reference and a claimed invention, however slight, invoke the question of obviousness, not anticipation. I have been advised that a claimed invention is not patentable under 35 U.S.C. §103 if it is obvious. A patent claim is unpatentable if the claimed invention would have been obvious to a person of ordinary skill in the field at the time the claimed invention was made. This means that even if all of the requirements of the claim cannot be found in a single prior art reference that would anticipate the claim, a person of ordinary skill in the relevant field who knew about all this prior art would have come up with the claimed invention.

14. I have further been advised that the ultimate conclusion of whether a claim is obvious should be based upon several factual determinations. That is, a determination of obviousness requires inquiries into: (1) the level of ordinary skill in the field; (2) the scope and content of the prior art; (3) what difference, if any,

existed between the claimed invention and the prior art; and (4) any secondary evidence bearing on obviousness.

15. I have been advised that, in determining the level of ordinary skill in the field that someone would have had at the time the claimed invention was made, I should consider: (1) the levels of education and experience of persons working in the field; (2) the types of problems encountered in the field; and (3) the sophistication of the technology.

16. I have also been advised that, in determining the scope and content of the prior art, in order to be considered as prior art, a reference must be reasonably related to the claimed invention of the patent. A reference is reasonably related if it is in the same field as the claimed invention or is from another field to which a person of ordinary skill in the field would look to solve a known problem.

17. I have been advised that a patent claim composed of several elements is not proved obvious merely by demonstrating that each of its elements was independently known in the prior art. In evaluating whether such a claim would have been obvious, I may consider whether there is a reason that would have prompted a person of ordinary skill in the field to combine the elements or concepts from the prior art in the same way as in the claimed invention.

18. I have been advised that any obviousness rationale for modifying or combining prior art must include a showing that a person of ordinary skill would have had a reasonable expectation of success.

19. I have been further advised that there is no single way to define the line between true inventiveness on the one hand (which is patentable) and the application of common sense and ordinary skill to solve a problem on the other hand (which is not patentable). For example, market forces or other design incentives may be what produced a change, rather than true inventiveness. I may consider whether the change was merely the predictable result of using prior art elements according to their known functions, or whether it was the result of true inventiveness. I may also consider whether there is some teaching or suggestion in the prior art to make the modification or combination of elements claimed in the patent. I may consider whether the innovation applies a known technique that had been used to improve a similar device or method in a similar way. I may also consider whether the claimed invention would have been obvious to try, meaning that the claimed innovation was one of a relatively small number of possible approaches to the problem with a reasonable expectation of success by those skilled in the art.

20. I understand that hindsight must not be used when opining on the patentability of claims of a patent. Therefore, my opinion will reflect the

understanding of a person of skill in the art as of the relevant patent application filing dates, which I have been instructed is August 16, 2019, for this matter, the earliest date to which applicants have claimed priority for the claims of the '794 patent.

IV. OVERVIEW

21. Long before the '794 patent, PacBio disclosed single-molecule real-time (SMRT) sequencing which allows for simultaneous sequencing of a large array of individual nucleic acid (e.g., DNA) molecules. *E.g.*, EX1023, 134-135. As shown in a video exemplifying SMRT sequencing such as described by Eid, sequence data is obtained by optically monitoring a polymerase as it incorporates fluorescently labeled nucleotides into a growing complementary chain. *Id.*; EX1044. As shown in screen shots from the video reproduced below, incorporation of a nucleotide is detected as a “pulse” of light. The optical pulse begins when the nucleotide (purple) binds to the active site of the polymerase (blue) and ends when the fluorophore is removed coincident with its incorporation into the growing chain. As shown, this process allows for sequential observation of nucleotide incorporation as the polymerase moves along the template molecule.



22. As reflected in the prior art literature, skilled artisans appreciated that the series of optical pulses provided much more information than just the sequence of the template molecule. *E.g.*, EX1006, 462; EX1007, 22:25-54; EX1009, Fig. 1. Specifically, skilled artisans knew that polymerase kinetics were reflected in the

pattern and timing of sequential pulses. Polymerase kinetic measures included the time the fluorescent nucleotide spent at the active site—termed pulse width (PW)—and the time between sequential nucleotide binding—termed interpulse duration (IPD). EX1006, Fig. 1; EX1007, Figs. 1 & 2. Skilled artisans appreciated (as was well-documented in the literature) that the structure of the template molecule, including its sequence and any epigenetic modifications (e.g., methylation), reproducibly affected PW and IPD. *E.g.*, EX1010, 130; EX1006, 462-464; EX1007, 4:66-5:3, 22:25-54, 62:29-31, 78:18-43, 79:2-5; EX1008, 1-2; EX1009, 124; EX1016, 3; *see also infra* Section V.B. Long before the '794 patent, sequence data and polymerase kinetic measures obtained in the course of SMRT sequencing were used to predict sites of epigenetic modification. As reflected in the prior art literature, this was one of the well-recognized advantages of SMRT sequencing relative to older sequencing technologies (e.g., Illumina sequencers). In this regard, skilled artisans developed models (e.g., machine learning) to analyze SMRT pulse data and characterized the effects of sequence and epigenetic modification on PW and IPD.

23. It is my opinion that the claims of the '794 patent are directed to using a model for detecting base modifications that takes as input these long-recognized sequence and kinetic parameters in SMRT pulse data. EX1021, 1-2; EX1008, 2; EX1009, 124, 126; EX1010, 130; *see also infra* Section V.C. I have reviewed

'794 patent in its entirety and it does not assert any novelty in the claimed SMRT pulse data parameters. *See generally* EX1001, 17:7-19:50; *see also infra* Section VII. This makes sense to me because, as discussed herein, the prior art discloses each of the claimed parameters—sequence context (identity and position), PW, and IPD—as conventionally generated in the course of SMRT sequencing and used to predict base modification. The '794 patent also does not assert any novelty in the generically claimed “model.” According to the specification, the claims broadly encompass statistical, machine learning, and other types of prior art models used to analyze genomic data including that obtained on the SMRT platform. *E.g.*, EX1001, 26:34-45, 29:39-43, 72:53-55, 77:39-67, 89:5-51; *see also infra* Section VII.

24. It is my understanding that the Patent Owner gained allowance of the claims of the '794 patent by representing that its “data structure” (nucleotide identity, position, PW, IPD) was a point of novelty. However, as discussed in detail herein, there is no novelty in the claimed data structure. The lack of novelty is demonstrated by, among other things, Patent Owner’s own briefing from district court. In arguing mapping to the claims, Patent Owner identifies a data structure (a BAM file) identically found in the published prior art long before the '794 patent. It is my opinion that the input data structure of the claims is provided by the BAM file which the prior art describes as generated in the normal course of SMRT

sequencing. *See infra* Section VIII. In my opinion, claims 1-19 of the '794 patent are obvious in view of prior art such as the Flusberg Publications, PacBio-BAM, Clement, Template Guide, and Tsai. I discuss the reasons for my opinion in detail below.

V. BACKGROUND IN THE ART

A. Third-Generation Sequencing

25. Prior to the '794 patent, a person of ordinary skill in the art would have been aware of multiple different commercial sequencing platforms. Persons of ordinary skill in the art often used the term “third-generation sequencing” to refer to single-molecule sequencing platforms where the base sequence is read directly from individual strands of DNA or RNA present in a sample of interest. EX1025, 2344. Third-generation sequencing was known and commercially available at least since 2010—that is, 9 years prior to the '794 patent’s earliest claimed priority date. EX1026, 1. This third-generation sequencing technology includes PacBio’s single-molecule real-time (“SMRT”) platform. A video (EX1044) briefly summarizing how SMRT sequencing works is available online at <https://www.pacb.com/technology/hifi-sequencing/how-it-works/>. The sequencing process described in the video corresponds to the disclosure of Eid et al. (EX1023).

26. One advantage that third-generation methods have over older next-generation sequencing (“NGS”) methods is that amplification of target nucleic acid

molecules is typically not required. EX1025, 2344. As such, target nucleic acid molecules may include bases with epigenetic modifications (e.g., methylation). EX1025, 2344-46. In contrast to older technologies, in third-generation sequencing such as PacBio's SMRT platform, a single DNA polymerase enzyme is observed in real time as it adds nucleotides to a growing strand. Because the enzyme processing rate was being recorded in third-generation sequencing, it was recognized that epigenetic base modifications (e.g., methylation) may be detected along with sequence data. *Id.* Because epigenetic modification plays a role in many critical biologic processes including gene expression, development, and cancer, *e.g.*, EX1007, 1:39-2:10, 15:26-16:59, there has been considerable interest in using third-generation sequencing platforms for modification detection. EX1006, 461-462; EX1007, Abstract; EX1008, Abstract; EX1009, Abstract; EX1010, Abstract; EX1015, Abstract; EX1016, 1; EX1022, Abstract; EX1025, 2345; EX1027, 18910; EX1035, Abstract; EX1038, Abstract. To this end, base modification detection methodologies have been developed for third-generation sequencing platforms such as PacBio's SMRT platform and the nanopore platform marketed by Oxford Nanopore Technologies ("ONT"). *E.g.*, EX1006; EX1007; EX1022.

B. SMRT Platform and Metrics Predictive of Modified Bases

27. Before the '794 patent, it was known that SMRT sequencing obtained sequence and polymerase kinetic data by optically monitoring a single polymerase as it incorporated nucleotides into a growing complementary strand. As illustrated below, SMRT sequencing generates a series of fluorescent pulses in which the color of each pulse indicates the identity of the incorporated nucleotide. The sequential pulses also provide kinetic parameters, including the width of the pulse (PW) and the time between pulses (IPD). EX1006, Fig. 1; *see also* EX1007, Figs. 1 & 2.

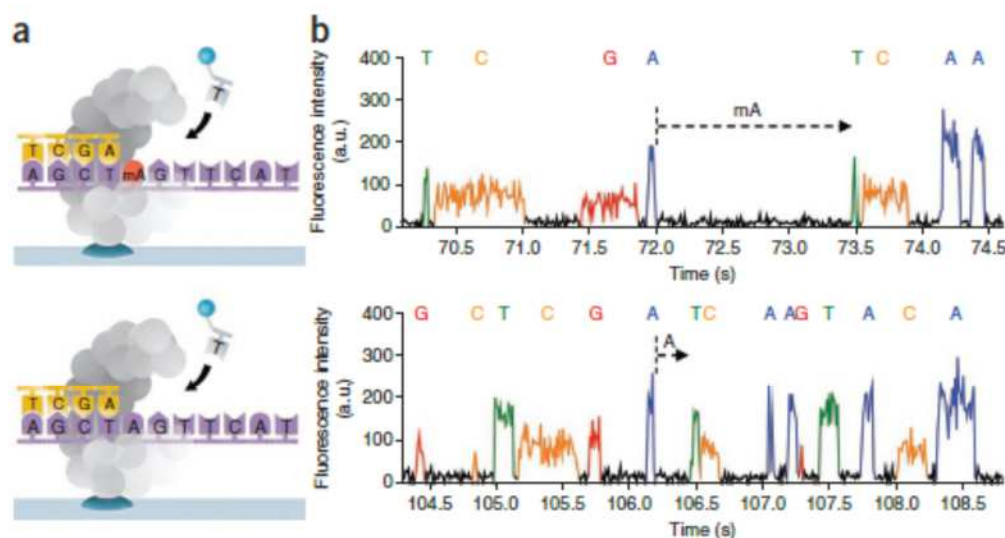
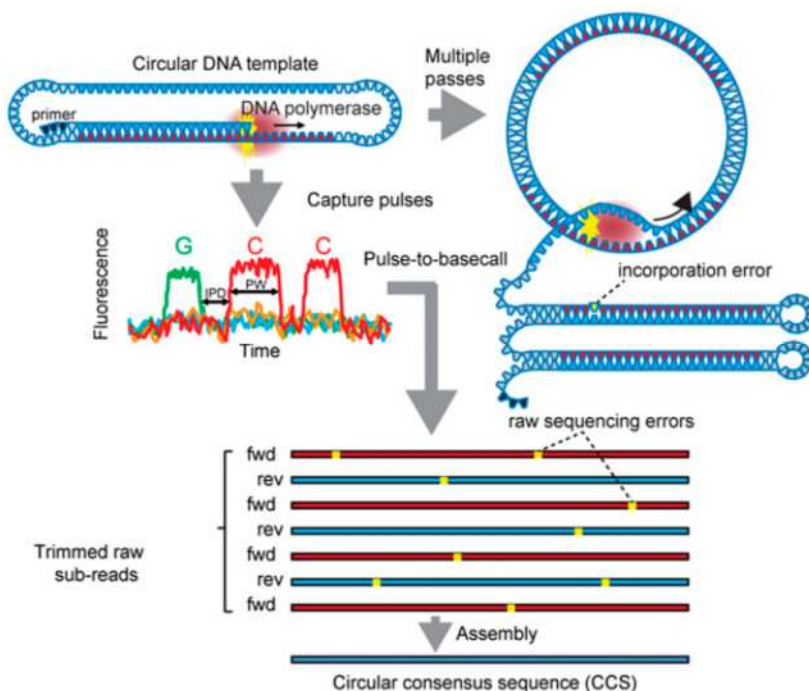


Figure 1 | Principle and corresponding example of detecting DNA methylation during SMRT sequencing. (a) Schematics of polymerase synthesis of DNA strands containing a methylated (top) or unmethylated (bottom) adenine. (b) Typical SMRT sequencing fluorescence traces for samples in a. Letters above the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

28. SMRT sequencing typically, although not necessarily, uses circular DNA templates—referred to as SMRTbells—which allow template molecules to be read multiple times by the polymerase as it repeatedly circles around the

SMRTbell. As shown below, multiple reads of the template would be used to generate a “circular consensus sequence” or “CCS” which minimized sporadic sequencing errors.



EX1009, Fig. 1. Multiple reads also facilitated “robust statistical analyses” of polymerase kinetic data, including pulse width (“PW”) and interpulse duration (“IPD”), which improved the accuracy of the data. EX1013, 194; *see also* EX1009, 127 (“[T]he PW and IPD of each pulse associated with the alignment were tabulated, allowing averages of kinetics to be taken for each alignment position across multiple molecules or within one molecule.”).

29. A person of ordinary skill in the art would have appreciated that, in addition to the sequence of the molecule, other useful metrics were generated by

SMRT sequencing. For example, pulse metrics such as the time between pulses (interpulse duration or IPD) and the width of the pulse (pulse width or PW) were known to provide valuable information regarding polymerase kinetics.

Fluorescence pulses in SMRT sequencing are characterized not only by their emission spectra but also by their duration and by the interval between successive pulses. These metrics, defined here as pulse width and interpulse duration (IPD), add valuable information about DNA polymerase kinetics. Pulse width is a function of all kinetic steps after nucleotide binding and up to fluorophore release, and IPD is determined by the kinetics of nucleotide binding and polymerase translocation.

EX1006, 462; *see also* EX1007, 22:25-54 (“While not all of these metrics are generally required for sequence determination, they add valuable information about the processing of a template, e.g., the kinetics of nucleotide incorporation and DNA polymerase processivity and other aspects of the reaction.”); EX1009, Fig. 1.

30. Prior to the '794 patent, persons of ordinary skill in the art appreciated that polymerase kinetics were dependent on sequence context. That SMRT polymerase kinetics was dependent on sequence context is well-documented in the art, including in the Schadt reference excerpted below.

Another interesting feature of the DNA polymerase kinetics in SMRT sequencing is that it is exquisitely sensitive to local sequence context. That is, 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, 130; *see also* EX1006, 462-463 (“As the only source of difference between these two loci in all three template types was the local sequence context ..., we conclude that, in general, the kinetic signatures of methylation will be sequence context-dependent.”); EX1007, 62:29-31 (“In particular, the local sequence context of the template strand can also influence and inhibit the activity of the polymerase along the template.”); EX1009, 124 (citing Flusberg) (“First, the variation in IPDs within the flanking sequences, even within short CGG-repeat or G-stretches, is much broader than within the CGG-repeat region itself, which is a reflection of the previously described sequence context effect on polymerase kinetics.”); EX1008, 1-2 (“The magnitude and extent of the kinetic signature is dependent on the type of the base modification and the local sequence context.”).

31. Persons of ordinary skill in the art also appreciated that that metrics associated with polymerase kinetics (e.g., IPD and PW) were used to detect modified bases (e.g., methylated bases). For example, Flusberg (EX1006) and Flusberg-338 (EX1007) describe using IPD and PW to identify modified bases. *E.g.*, EX1006, 463 (“In fact, because each modification had a unique IPD and pulse width signature in a given context ..., this approach opens up the intriguing possibility of directly distinguishing between cytosine, mC and hmC during real-

time sequencing.”),¹ Fig. 3; EX1007, 22:25-54 (“For example, the presence of modified bases in a template nucleic acid has been shown to change the width of a pulse and/or the interpulse duration (IPD)....”).

32. Clark further explains that PW and IPD are the principal metrics used to identify modified bases.

The length of time that the polymerase retains a nucleotide bound in its active site (pulse width, PW), and the time interval between successive nucleotide bound states (interpulse duration, IPD) are the principal pulse metrics used in the analysis to ascertain that the polymerase kinetics was altered in the modification-containing template when compared to an unmodified control template.

EX1008, 1-2 (citing Flusberg). Clark demonstrates that these same metrics may be used to identify various types of damaged bases (e.g., 8-oxoguanine, 8-oxoadenine, O6-methylguanine, 1-methyladenine, O4-methylthymine, 5-hydroxycytosine, 5-hydroxyuracil, 5-hydroxymethyluracil, or thymine dimers). EX1008, Abstract. *See generally* EX1008, 1-9.

¹ The abbreviations “mC” and “hmC” refer to methylcytosine and hydroxymethylcytosine. Similar shorthand nomenclature is used for methylated (e.g., mA) or oxidized adenosine (e.g., oxo-A).

33. The presence of a modified base was known to alter polymerase kinetics at nearby bases in a manner diagnostic of the type of base modification, further illustrating the recognized significance of local sequence context. This concept is explained in a white paper published by PacBio, among many other references.

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

EX1016, 3; *see also* EX1010, 130 (“The strong dependence of IPD on local sequence context, combined with the fact that modifications at one base can affect the IPD at nearby bases, support that modifications to bases change the local sequence context in ways that alter how the local sequence (the 10-mer) interacts with the enzyme, and these interactions can be nonlinear, with changes in one base affecting the kinetics of the enzyme at nearby bases.”); EX1006, 462-464 (“For all three methylation types, there was a clear excursion in polymerase synthesis kinetics in the vicinity of the methylated bases (Fig. 2). The methylated base is in contact with the polymerase for several bases before and after occupying the polymerase active site.”), Fig. 2; EX1007, 4:66-5:3, 22:25-54 (“For example, the

presence of certain modifications alters not only the processing of the template at the site of the modification, but also the processing of the template upstream and/or downstream of the modification.”), 78:18-43, 79:2-5, Figs. 18-19; EX1008, 2 (“The magnitude and extent of the kinetic signature is dependent on the type of the base modification and the local sequence context. This results in distinct, more complex kinetic signatures that can be used to discriminate between different chemical base modifications.”).

34. Prior to the ’794 patent, persons of ordinary skill in the art appreciated that SMRT sequencing generated values for the identity of each nucleotide and its position along with other pulse metrics. These values were stored together in a data structure referred to as a “BAM file.” EX1011, 1. PacBio’s BAM file included sequence information and PacBio-specific metrics such as IPD and PW. EX1011, 1-2, 4-5, 8. PacBio-BAM file documentation published prior to the ’794 patent explains that PW and IPD values for each nucleotide were stored as either an 8- or 16-bit integer array. This documentation describes PacBio BAM files as “fully compatible with the BAM specification,” which specifies that a BAM file includes the identity of each nucleotide, in sequence, as an array of 4-bit integers. EX1011, 1; *see* EX1012, 15.

35. To summarize, it was well-known and documented in the prior art that IPD and PW values in the vicinity of a base of interest are predictive of not only

the modification state of a base but even the particular type of modification.

Because the sequence context (e.g., identity of the bases in the vicinity of a modified base) strongly affects IPD and PW, it was further appreciated that context must be controlled for or explicitly evaluated when predicting base modifications.

C. Predictive Models

36. Prior to the '794 patent, metrics predictive of modification state were known and it was also known models could be trained with these metrics and used to identify modified bases. The inventors of the '794 patent were apparently of the same opinion as the patent includes discussion of prior art methods using models such as deep learning, hierarchical, conditional random field, and principal component analysis to identify modified bases in data generated by SMRT and ONT platforms. EX1001, 17:32-35, 19:2-11, 19:14-18, 35:21-37.

37. The prior art taught multiple different models using third-generation sequencing data as input to detect base modifications. One example is described in PacBio's kineticsTools manual. The kineticsTools manual is a file available from GitHub, as archived by the Internet Archive's Wayback Machine at <http://web.archive.org/web/20170813085437/https://github.com/PacificBiosciences/kineticsTools/blob/master/doc/manual.rst> (EX1021, archived Aug. 13, 2017). The kineticsTools manual describes a model for detecting modified bases that receives a PacBio BAM file as an input to the model and uses the information in

that file, including kinetic data (IPD values) and a window over a sequence of nucleotides to predict whether particular bases are modified by the presence of a methylation on a nucleotide of interest located within the window. EX1021, 1-2 (“The bounds of the relevant context window correspond to the window of DNA in contact with the polymerase, as seen in DNA/polymerase crystal structures. To simplify the process of finding DNA modifications with PacBio data, the tool includes a pre-trained lookup table mapping 12-mer DNA sequences to mean IPDs observed in C2 chemistry.”), 2 (“Basic use with BAM input, GFF output.... A standard PacBio alignment file - either AlignmentSet XML or BAM - containing alignments and IPD information supplies the kinetic data used to perform modification detection.”).

38. Other references further illustrate models which receive as input a combination of kinetic parameters (including IPD and PW) and the local sequence context over a window of nucleotides. For example, Schadt teaches that both PW and IPD are affected by modifications, with IPD being more strongly affected, and further teaches that polymerase kinetics “is exquisitely sensitive to local sequence.” EX1010, 130 (citing Flusberg, EX1006). Indeed, Schadt teaches that the nucleotide identities over a 10-nucleotide window including a nucleotide of interest produces “>80% of the variance in the IPD measures.” *Id.*; *see also* EX1008, 2 (“The magnitude and extent of the kinetic signature is dependent on the

type of the base modification and the local sequence context.”); EX1009, 124, 126 (confirming effect of local context on polymerase kinetics).

39. Various third-generation sequencing technologies known and commercially available at the time used a combination of kinetic metrics and sequence context as inputs to a model for predicting modifications. For example, Liu et al. teach using a trained neural network (“DeepMod”) to determine methylation from ONT sequencing data. EX1022, Abstract, 3. According to Liu, the input for DeepMod is a FAST5 file—the standard file format for ONT data—which includes raw signal and sequence data.

In DeepMod, the input is a reference genome and FAST5 files generated by Nanopore sequencers with raw signals and base calls, and the output includes modification prediction for bases of interest in a long read and modification summary for genomic positions of interest in a reference genome ...

EX1022, 8. The ’794 patent confirms that Liu had shown it was possible to “detect methylation states with the use of a deep learning method,” differing merely in that it used measurements relevant to ONT sequencing rather than the fluorescence parameters relevant to SMRT sequencing (for example, Liu uses electric pulse characteristics in combination with basecalls instead of fluorescence pulse characteristics in combination with basecalls). *See* EX1001, 17:7-63. With regard to the particular structure of Liu’s model, Liu describes using a neural network that

receives input in the form of a data structure containing both kinetic features and nucleotide identities for each nucleotide in a window centered around a nucleotide of interest, with the output indicating whether the base in the center of the window is modified. EX1022, 8-9, Fig. 1 (reproduced below).

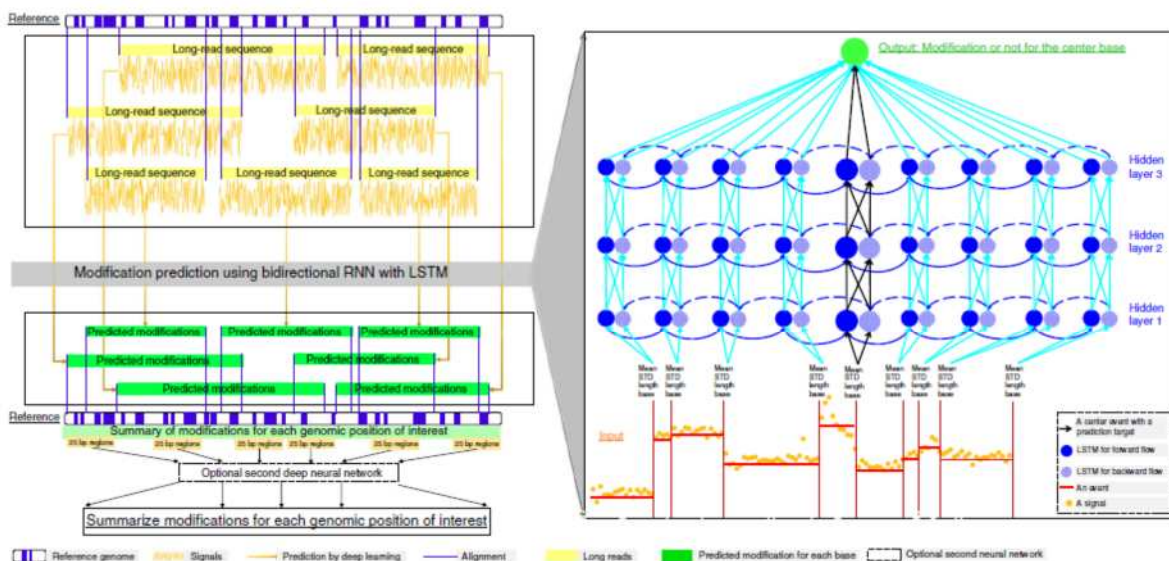


Fig. 1 The flowchart of DeepMod. RNN: recurrent neural network; LSTM: long short-term memory. Several long reads (in yellow) were shown for demonstration purposes with their alignment and Nanopore signals (yellow lines). To view LSTM RNN for modification prediction, nine LSTM cells adjacent to the center position (with arrows in black) were shown

40. Another example of prior art modification algorithms is that described by Liu, et. al for detection of methylated adenosine (m6A) bases in RNA. EX1038, Title (“Accurate detection of m6A RNA modifications in native RNA sequences.”). Liu describes using “Principal Component Analysis (PCA)” and “Support Vector Machines (SVM)” to analyze ONT data. EX1038, 3; *see also* EX1001, 29:43-51 (“Other algorithms could also be used in addition or instead, for example, but not limited to...support vector machine (SVM)”), claim 11 (“principal component analysis”). According to Liu, use of any one parameter

(e.g., base quality, mismatch, and deletion frequency) led to low predication accuracy. EX1038, 4. However, inclusion of all three parameters together “increased the accuracy of the model.” EX1038, 4. Liu explains that its data is publicly available and may be used to “train different machine learning algorithms” which may lead to “improved methods to detect RNA modifications in the future.” EX1038, 4.

41. A person of ordinary skill in the art would have known that other models, such as convolutional neural networks, were also used for prediction of methylation in ONT sequencing data. The ’794 patent asserts that because “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.” EX1001, 17:42-48. However, this statement is misleading. The specific measured kinetic parameters for ONT systems differ from those of PacBio—for example, ONT measures voltage characteristics and time between translocations, whereas PacBio measures pulse widths and interpulse duration. Yet skilled artisans appreciated that the same methods used to model modifications for ONT were applicable to PacBio SMRT sequencing data by substituting the specific kinetic parameters applicable to each system. *See, e.g.,*

EX1022, Abstract, 3, 8-9, Fig. 1. In one example, a paper by Liu et al. drew parallels between use of PacBio and ONT’s kinetic parameters for detecting modifications, noting that in each system, one can detect “modifications by measuring the kinetics,” with only the specific kinetic parameters varying—as ONT’s kinetic parameters involve “disruptions in the pore current” when nucleic acid is pulled through a nanopore, whereas PacBio measures “kinetics of the reverse transcriptase” as it encounters the nucleic acid molecule. EX1038, 2. The use of similar techniques of input and training in different third-generation systems underscores the practical reality that, regardless of sequencing platform, a person of ordinary skill in the art would have predicted methylation using models by training with features associated with methylation for the platform in question.

D. Deep Learning and Computing Power

42. In the years since PacBio released SMRT sequencing, the commercial viability of machine learning techniques had been further enhanced due at least in part to advancements in computing power and deep learning methods (e.g., neural networks). As explained in a 2019 NIH report, “computing power has increased by orders of magnitude” in the past decade. EX1028, 787. Additionally, cloud-based services allowed access to high-performance computing without the necessity of investing and maintaining dedicated machines. EX1028, 787 (“Cloud computing is typically more cost-effective when bursts of high-performance computing are

needed, while sustained computing is typically more cost-effective on dedicated machines in local data centers.”). These improvements coincided with rapid development of deep learning networks which required high-performance computing to train and run. EX1031, 2 (“According to OpenAI, the demand for compute by deep learning networks has been doubling every 3.5 months since 2012.”).

43. While machine learning was known for decades, these technologies saw increasingly widespread implementation in genomic applications and elsewhere, in conjunction with increased accessibility via open-source resources. EX1028, 782. These developments further underscored the feasibility of commercial implementation of such techniques to a wide array of problems, including those presented in genomics. By 2019 numerous deep learning techniques had been developed, with many available as open-source software. *E.g.*, EX1017, 16 (collecting open-source “Deep learning resources” for genomic applications). In fact, the inventors of the ’794 patent acknowledged in a subsequent paper that their convolutional neural network model was implemented using an open-source deep learning framework. *See* EX1030, 10. In particular, the inventors stated that the “program for the CNN model was implemented on the basis of the Keras deep learning framework.” *Id.* Keras is an open-source platform providing high-level abstraction for developing deep learning models, making it

possible for users with basic skills in the Python programming language to build and train commonly used neural networks. *See* EX1017, 16-17 (discussing the Keras open-source platform); EX1035, 3 (“the proposed model is implemented by Keras framework”).

44. Prior to the ’794 patent, deep learning networks had been applied to a number of problems in the field of genomics including those involving third-generation sequencing. EX1017, 12-13 (“Machine learning techniques have been extensively used in genomics research.”), 14 (discussing application of deep learning to methylation), 17 (“Interactive tutorial to build a convolutional neural network....”); EX1022, Abstract (“DeepMod performs well for genome-scale detection of DNA modifications and will facilitate epigenetic analysis on diverse species.”), 10 (“DeepMod is publicly available”); EX1029, 5 (“More recent works apply CNNs directly on the raw DNA sequence, without the need to define features a priori.”); EX1035, 2 (“The proposed method is a simple and efficient architecture for 2’-O-methylation sites.”); EX1038, 3-4.

45. In sum, by August 16, 2019, deep learning represented an increasingly widespread approach to analyzing complex datasets such as those obtained from third-generation sequencing platforms. *See, e.g.*, EX1032, 1-7 (August 12, 2019 PacBio paper describing application of Google’s DeepVariant network to SMRT data); EX1036, 1 (“DeepVariant is a deep learning-based variant caller that takes

aligned reads (in BAM or CRAM format), produces pileup image tensors from them, classifies each tensor using a convolutional neural network, and finally reports the results in a standard VCF or gVCF file.”).

VI. PRIOR ART

A. The Flusberg Publications (Flusberg and Flusberg-338)

46. “Direct detection of DNA methylation during single-molecule, real-time sequencing” by Flusberg et al. published online in Nature Methods on May 9, 2010 along with Supplemental Information (“Flusberg,” EX1006). I understand that Flusberg may be considered in evaluating the patentability of the claims of the ’794 patent.

47. U.S. Patent No. 9,175,338 to Flusberg, et al is entitled “Methods for identifying nucleic acid modifications” (“Flusberg-338,” EX1007). Flusberg-338 incorporates Flusberg by reference in its entirety. EX1007, 77:30-34, 78:58-66. Flusberg-338 was filed on November 12, 2010 and issued on November 3, 2015. I understand that Flusberg-338 may be considered in evaluating the patentability of the claims of the ’794 patent.

48. I will refer to Flusberg and Flusberg-338, collectively, as “the Flusberg Publications.”

49. I discuss the Flusberg Publications along with other references in Section V. In brief, the Flusberg Publications describe methods for detecting base

modifications, such as methylated cytosine, that consider sequence context and polymerase kinetics. The Flusberg Publications disclose that sequence context and base modifications alter polymerase kinetics in a manner that allows modified bases to be detected and even categorized (e.g., 5mC or 5hmC). The Flusberg Publications further disclose that polymerase kinetics is reflected in the PW and IPD metrics which are obtained along with base identity from SMRT sequencing. *E.g.*, EX1006, 462-463, 476; EX1007, 22:25-54, 62:29-38; *see also supra* section V.B.

50. The Flusberg Publications further describe models (e.g., statistical, machine learning) for identifying modified bases in template molecules using data obtained from SMRT sequencing. *E.g.*, EX1006, 462-463, Fig. 3; EX1007, 61:6-69:51. For example, the Flusberg Publications disclose using models to identify modified bases by analyzing IPD and PW values obtained for a potentially modified nucleic acid molecule in comparison to values obtained for an unmodified control of the same sequence. EX1006, 463-465, Figs. 1 & 3; EX1007, 78:18-43. Flusberg-338 also discloses training machine learning models with data (e.g., sequence context, IPD, and PW) obtained from sample molecules having a known modification state (e.g., unmodified) and using the trained model to identify modified bases. EX1006, 463-465; EX1007, 64:29-48, 66:46-67:36.

B. PacBio-BAM

51. PacBio published on its file format website documentation describing the BAM files used to store data obtained from SMRT sequencing (“PacBio-BAM,” EX1011). This website was archived by the Internet Archive on September 19, 2017, at <http://web.archive.org/web/20170916112547/http://pacbiofileformats.readthedocs.io/en/5.0/BAM.html>. I understand that PacBio-BAM may be considered in evaluating the patentability of the claims of the ’794 patent.

52. PacBio-BAM describes the content and exemplary uses of the BAM file generated in the normal course of SMRT sequencing. A person of ordinary skill in the art would have appreciated that a BAM file is an industry-standard file format. PacBio-BAM describes the PacBio BAM file as storing “raw or aligned sequence reads” along with SMRT-specific metrics such as PW and IPD. EX1011, 1. PacBio-BAM further indicates that pulse features (e.g., PW and IPD) are included and may be used for downstream analyses including base modification detection. EX1011, 8 (indicating the BAM file may be used for “base modification detection”).

53. PacBio-BAM was also provided to the public via other known sources. For example, a paper by Lau et al. published in 2016 cited to “BAM format specification for PacBio,” at the following GitHub site: <https://github.com/PacificBiosciences/PacBioFileFormats/blob/3.0/BAM.rst>. GitHub is website providing online services for software development and version control. I am

familiar with GitHub as a user and a contributor, and I have uploaded source code and related contributions to GitHub continually over the past decade. GitHub allows developers to store and manage their code, as well as associated documentation, and make them available to the public. GitHub tracks the date of upload of files as well as tracking the date any edits to an uploaded file are made.

54. I accessed the URL provided by Lau et al., which linked to a GitHub page providing the 3.0 version of the Pacific Biosciences BAM format specification. EX1042. The page also included a menu allowing me to access the 5.0 version of the Pacific Biosciences BAM format specification. EX1043. I inspected the 5.0 version, and its contents are the same as those in PacBio-BAM (EX1011). The 3.0 version also contains the same relevant disclosures as the 5.0 version, with only minor differences not material to my conclusions in this declaration. The GitHub page I visited indicates that the 5.0 version of the PacBio BAM specification (<https://github.com/PacificBiosciences/PacBioFileFormats/commits/5.0/index.rst>) was last modified on March 7, 2017, and the 3.0 version of the PacBio BAM specification (<https://github.com/PacificBiosciences/PacBioFileFormats/commits/5.0/index.rst>) was last modified on March 28, 2016. Based on my experience with GitHub, these dates indicate that the 5.0 and 3.0 versions were publicly available from GitHub in their current form since March 7, 2017, and March 28, 2016, respectively.

C. Clement

55. U.S. Patent Publication No. 2016/0326593 to Clement et al. (“Clement,” EX1018) is entitled “Compositions and Methods for Diagnosing, Evaluating and Treating Cancer.” Clement was filed on May 23, 2016 and published on November 10, 2016. I understand that Clement may be considered in evaluating the patentability of the claims of the ’794 patent.

56. Clement discloses “methods of determining a cancer treatment prognosis” based on “evaluating epigenetic and genetic changes within a tumor sample.” EX1018, Abstract. Clement describes classifying disease states according to methylation pattern detected from a sample. EX1018, ¶¶9, 78, 198.

D. Template Guide

57. PacBio published the Template Guide (EX1019) on their website, which was archived by the Internet Archive on February 7, 2016, at http://web.archive.org/web/20160207131429if_/http://www.pacb.com:80/wp-content/uploads/2015/09/Guide-Pacific-Biosciences-Template-Preparation-and-Sequencing.pdf. I understand that the Template Guide may be considered in evaluating the patentability of the claims of the ’794 patent.

58. The Template Guide describes recommended preparation methods for DNA templates libraries for sequencing on the SMRT platform. It further describes

selecting a preferred size range of templates as well as removing templates below a defined size. EX1019, 24-28, 35.

E. Tsai

59. U.S. Patent Publication No. 2016/0208241 to Tsai et al. (“Tsai,” EX1020) is entitled “Compositions and Methods for Enrichment of Nucleic Acids.” Tsai was filed on January 14, 2016 and published on July 21, 2016. I understand that Tsai may be considered in evaluating the patentability of the claims of the ’794 patent.

60. Tsai is a publication of a patent application assigned to PacBio. Tsai describes an amplification-free enrichment technique using the CRISPR-Cas9 system to generate SMRT sequencing libraries. EX1020, ¶¶6-7, 65-66, Fig. 8. Tsai teaches sequencing “enriched nucleic acids” using PacBio SMRT sequencing, teaching that “[p]referably the sequencing method is capable of detecting base modifications during a sequencing reaction, e.g., by monitoring the kinetics of the sequencing reaction.” EX1020, ¶23. Tsai further cites the international application (PCT publication) of the Flusberg-338 (EX1024) as providing “[m]ore information on detection of modifications during sequencing reactions.” *Id.*

VII. THE ’794 PATENT

61. The ’794 patent is entitled “Determination of Base Modifications of Nucleic Acids.” EX1001, cover (54). A person of ordinary skill in the art would

understand the alleged invention of the '794 patent to be the product of following instructions that were long ago described in the prior art such as, using sequence context and polymerase metrics (e.g., IPD and PW) obtained from SMRT to detect base modifications. EX1001, 2:54-64, 19:54-60.

62. Claim 1 is verbose but requires nothing beyond practicing art known methods for identifying modified bases in sample nucleic acids. Claim 1 requires acquiring data generated in the normal course of SMRT sequencing, storing it in an input data structure (e.g., BAM file), and then inputting the stored data into a trained model to determine modified bases. The recited properties which are acquired, stored, and input into a model are sequence context (“identity” and “position”) and metrics associated with polymerase kinetics (“width of the pulse” and “interpulse duration”).

63. The '794 patent provides a very narrow reading of the prior art. *See generally* EX1001, 17:7-19:50. But, even still, it acknowledges that nearly every aspect of the claimed subject matter was known in the art. For example, the '794 patent acknowledges that third-generation sequencing platforms, including PacBio's SMRT platform, have long been used to detect modified bases. EX1001, 13:7-17 (citing Schrieber (EX1027) and Flusberg (EX1006)) (“[P]rocesses known to those skilled in the art can be used to interrogate the methylation status of DNA molecules, including, but not limited to ..., single molecule sequencing using a

platform sensitive to the methylation status (e.g. nanopore sequencing ... and by single molecule, real-time sequencing (e.g. that from Pacific Biosciences) ...).”).

64. The ’794 patent acknowledges that it was known in the art that sequence context and polymerase kinetics were used to predict base modification. EX1001, 19:2-13 (discussing a method that uses IPD and sequence context), 19:38-40 (“...these previous attempts of utilizing IPD only or with combination of sequence information in the neighboring nucleotides...”). *See generally* EX1001, 17:7-19:50. The ’794 patent also credits Flusberg for demonstrating “polymerase kinetics would be affected by methylation states in the sequences.” EX1001, 17:64-65. The ’794 patent acknowledges that Flusberg analyzed polymerase kinetics, as reflected in IPD and PW, to detect various base modifications. EX1001, 18:19-30; *see also* EX1006, 462 (“These metrics, defined here as pulse width and interpulse duration (IPD), add valuable information about DNA polymerase kinetics.”).

65. The ’794 patent characterizes Flusberg as describing pulse width as not significantly affected by base modifications. EX1001, 18:19-30. I disagree. Flusberg expressly describes base modifications as having a significant, measurable effect on pulse width. *See* EX1006, 463 (“[E]ach modification had a unique IPD and pulse width signature in a given context.”).

66. The '794 patent indicates that a broad array of art-known models are within the scope of the invention. EX1001, 26:34-45 (“[A]ny model or data analysis framework can be developed.”), 72:53-55 (“In other embodiments, other neural network models, deep learning algorithms, artificial intelligence, and/or machine learning algorithms could be used.”), 89:5-51. For example, the '794 patent indicates that models within the scope of the invention include conditional random field, deep learning, hierarchical models, and principal component analysis. EX1001, 29:39-43, 77:39-67, 89:25-51, claim 11. The '794 patent acknowledges that these same models had been used previously for detecting modified bases in third-generation sequencing data. EX1001, 17:32-37 (“enabled one to detect methylation states with the use of a deep learning method”), 19:2-8 (“Feng et al. used a hierarchical model...”), 19:14-18 (“Schadt et al. developed a statistical method, called conditional random field, ...”), 35:21-37 (“principal component analysis”). The '794 patent also describes input data structures that may contain values for relevant properties used by models. *See* EX1001, 87:42-88:3, 90:20-91:31, Figs. 102, 103 (indicating the input data structure “correspond[s] to a window of nucleotides sequenced in a sample nucleic acid molecule” and contains values for metrics obtained from SMRT sequencing (e.g., nucleotide identity, position, PW, IPD). However, a person of ordinary skill in the art would have understood that the various types of models that the '794 patent

acknowledges were used in the prior art for methylation detection would also have been trained using input data structures containing relevant properties correlated with methylation status in order to learn to accurately detect methylation.

67. The '794 patent suggests that its novel contribution is the combined use of multiple parameters (sequence context, IPD, and PW) that it acknowledges were known for detection of base modifications.

We reasoned that if we can use a combination of multiple metrics, for example, concurrently making use of features including upstream and downstream sequence context, strand information, IPD, pulse widths as well as pulse strength, we might be able to achieve the accurate measurement of base modifications (e.g. mC detection) at single-base resolution.

EX1001, 19:54-60.

68. However, it is well-documented in the literature—and acknowledged in the '794 patent—that each of these metrics was predictive of modification state. Furthermore, values for each of these metrics were routinely generated in the ordinary course of SMRT sequencing, included in PacBio's generated BAM file, and users were instructed to use kinetic data in the BAM file for modification detection. As I explain in detail herein, making use of multiple metrics known to be predictive of modification state which were routinely obtained from SMRT sequencing was known or, at a minimum, would have been obvious to a person of ordinary skill in the art in view of the prior art.

VIII. PACBIO-BAM AND THE TAKE2 INFRINGEMENT CONTENTIONS

69. I have been asked to review Take2's infringement contentions regarding the '794 patent (EX1014), and in particular to compare the citations therein to the Pacific Biosciences BAM Specification, both as it appeared in February 2022 (PacBio-BAM[2022], EX1033) and as it appeared in 2016 (PacBio-BAM[2016], EX1011).² As I discuss below, Take2's infringement contentions map the data structure elements of claim 1 to disclosure of PacBio-BAM[2022] that also appears in PacBio-BAM[2016].

70. On page 15, the Infringement Contentions cite to section 2.9 of PacBio-BAM[2022], "Use of read tags for per-read information." EX1014, 15 (citing EX1033, 10). The same table is present in PacBio-BAM[2016]. EX1011, 1-2.

71. On page 25, the Infringement Contentions cite to two sections of PacBio-BAM[2022]. EX1014, 25 (citing EX1033, 8, 9). The second cited portion is found verbatim in PacBio-BAM[2016]. EX1011, 1. The first cited portion discusses the fact that reads may be aligned or unaligned; while it is not present

² In this section, I refer to the different versions of PacBio-BAM with their dates in brackets. In the remainder of this declaration, I refer to EX1011 as PacBio-BAM.

verbatim, PacBio-BAM[2016] also discloses that reads may be aligned or unaligned. EX1011, 1.

72. The portion of the PacBio-BAM[2022] cited on page 27 of the Infringement Contentions is found verbatim in PacBio-BAM[2016]. *See* EX1014, 27; EX1011, 1; EX1033, 5.

73. The portion of PacBio-BAM[2022] cited on page 30 of the Infringement Contentions is found verbatim in PacBio-BAM[2016], including an identical figure showing alignment of a read as a query to a target sequence. *See* EX1014, 30; EX1011, 1-2; EX1033, 10.

IX. OVERVIEW OF THE LEVEL OF SKILL

74. I have been advised that “a person of ordinary skill in the relevant field” (or “person of ordinary skill in the art,” or “POSA”) is a hypothetical person to whom one could assign a routine task with reasonable confidence that the task would be successfully carried out. I have been advised further that, in determining the level of ordinary skill in the art that someone would have had at the time the claimed invention was made, I should consider: (1) the levels of education and experience of persons working in the field; (2) the types of problems encountered in the field; and (3) the sophistication of the technology.

75. In my opinion, a person of ordinary skill in the relevant field as of August 16, 2019, would typically have M.S. or Ph.D. degree with 2-4 years of

research experience in genomics, bioinformatics, computational biology, or related fields, or could have less education but significant professional experience in one or more of these fields.

X. CLAIM CONSTRUCTION

76. I have been advised that the claims terms of the '794 patent claim are to be given the plain and ordinary meaning that the terms would have to a person of ordinary skill in the art in question at the relevant priority date. I understand that the analysis focuses on how the patentee used the claim term in the claims, specification, and prosecution history. I also understand that dictionaries or other extrinsic sources may assist in determining the plain and ordinary meaning but cannot override a meaning that is unambiguous from the intrinsic evidence. I have followed these principles in my analysis throughout this declaration.

77. I also understand that inventors may act as their own lexicographers and provide definitions for terms. For example, claims 3-6, 8, 9, and 11, contain terms expressly defined in the '794 patent. EX1001, 14:65-15:5 ("classification"), 15:6-24 ("cutoff"), 16:34-40 ("machine learning models"). For these terms, I have applied the definitions provided in the specification.

XI. GROUND 1 – CLAIMS 1-2, 7, 10-12, 14-15, AND 17-19 ARE OBVIOUS AS ILLUSTRATED BY THE FLUSBERG PUBLICATIONS AND PACBIO-BAM

78. As discussed in detail below, the Flusberg Publications in view of PacBio-BAM render claims 1-2, 7, 10-12, 14-15, and 17-19 obvious. Below, I note

specific prior art disclosure that satisfies each element of each challenged claim.

After each claim element, I provide a chart identifying certain exemplary prior art disclosure that is relevant to that element.

79. The Flusberg Publications disclose a method for determining whether a nucleotide is modified, such as by being methylated. EX1006, Abstract; EX1007, Abstract. The Flusberg Publications disclose that nucleic acid molecules are sequenced using PacBio SMRT sequencing, generating pulse traces in which each nucleotide produces a pulse from which characteristic features may be derived. EX1006, 462-63, Fig. 1; EX1007, 23:14-22, 64:29-43, 66:46-67:4, Figs. 1 & 2. These features include the identities of each nucleotide in sequence, as well as parameters including PW and IPD for each pulse. *See* EX1006, 462-63, Fig. 1; EX1007, 23:14-22, 64:29-65:20, 66:46-67:4.

80. The Flusberg Publications teach that pulse features (including pulse width and interpulse duration) may be tabulated for a window of nucleotides and used in a trained model to determine whether a target nucleotide is methylated. *E.g.*, EX1006, 463, 465-466, Fig. 3; EX1007, 66:46-67:36. For example, Flusberg teaches using a principal component analysis model that takes a data structure including IPD and pulse width signals over a window of nucleotides as input and outputs a signal discriminating between different modification states (such as unmodified, methylated, and hydroxymethylated) at a given nucleotide. EX1006,

463. Flusberg-338 similarly describes analytic models for determining methylation state that “can be implemented with many standard statistical classification algorithms” and trained with pulse data (e.g., context, PW, and IPD) over a window of nucleotides. EX1007, 66:46-68:48.

81. Regarding the claims’ recitation of an input data structure including the properties of nucleotide identity and position, pulse width, and IPD in an input data structure, this is taught by PacBio-BAM, which teaches the standard data structure in which PacBio SMRT sequencers store data. EX1011, 1-2. These data structures store, *inter alia*, identity, position, pulse width, and IPD values for each nucleotide in a sequence, such as a window of nucleotides around a target nucleotide for which a modification state is to be determined, as taught by the Flusberg Publications. EX1011, 1-2, 4-5, 8; EX1012, 15.

A. Claim 1

82. A person of ordinary skill in the art would have had good reason to combine the Flusberg Publications with PacBio-BAM to achieve what is recited in the claims with a reasonable expectation of success. The Flusberg Publications disclose detecting modifications using PacBio SMRT sequencing. PacBio-BAM teaches a BAM file containing values for all claimed properties that was generated in the normal course of SMRT sequencing. As PacBio-BAM shows, BAM files are generated by the SMRT sequencing platform which contain data for the identity

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

83. It was well-known that values for the claimed SMRT sequencing properties were routinely generated and known as useful in base modification detection. Specifically, skilled artisans appreciated that the sequence context (including the identity and positions of the bases in the vicinity of a modified base) strongly affects polymerase kinetics, and as a result, IPD and PW values. *E.g.*, EX1006, 462-463; EX1007, 62:29-31; EX1008, 1-2; EX1009, 124; EX1010, 130; EX1016, 2. It was further appreciated that polymerase kinetics was also dependent on the presence of base modifications and that different modification types had unique PW and IPD signatures. *E.g.*, EX1006, 462, Fig. 2; EX1007, 4:66-5:3, 22:45-54, 78:18-43, 79:2-5, Figs. 18-19; EX1008, 2; EX1010, 130; EX1016, 3.

84. Because of the well-documented interdependence of local sequence context and base modifications on PW and IPD, skilled artisans appreciated that local sequence effects had to be considered when predicting base modifications. Flusberg expressly teaches as much, concluding that, “in general, the kinetic signatures of methylation will be sequence context-dependent.” EX1006, 462-463. Other prior art came to the same conclusion; for example, Schadt et al. found that sequence context very strongly affected polymerase kinetics, noting that the context over a window running “seven bases upstream of and two bases downstream from the incorporation site explain >80% of the variance in the IPD measures.” EX1010, 130; *see also supra* Sections V.B-C. A person of ordinary skill in the art would have appreciated that, because these different properties were interrelated, they should be considered together when classifying modifications across varying sequence contexts. For example, a person of ordinary skill in the art would understand that, based on the prior art’s recognition that sequence context accounts for >80% of the variance in IPD, any variation in that context must be accounted for when detecting modifications using kinetic parameters like IPD, or else the sequence context will act as an uncontrolled confounder that drowns the signal for methylation detection in noise. A person of ordinary skill in the art would thus have found natural to use the claimed properties together to detect modifications in target nucleotides and doing so would represent nothing beyond

the predictable use of known variables according to their well-characterized role in base modification detection.

85. The scientific literature prior to August 16, 2019, further corroborates the obviousness of considering data for nucleotide identity, position, PW, and IPD together. For example, even prior to the adoption of the BAM file format by PacBio, Loomis describes generating a file containing SMRT sequencing data for analysis of polymerase kinetics. EX1009, 127. This data file included the base identities in order linked with pulse parameters such as PW and IPD. EX1009, 127. Loomis describes performing a statistical analysis on this data which showed “differences in the rates of polymerization” across different regions, and even individual bases, of a template molecule. EX1009, 124, 127. This analysis confirmed the interplay between kinetic signatures of modifications, such as IPD and PW, and sequence context, as differences in kinetics were observed based on the local identities and positions of nucleotides around a nucleotide of interest. *See* EX1009, 124 (“The principal outcome measure for the time-domain measurements is the interpulse distance (IPD) The other outcome measure is the pulse width (PW)”), (“This observation likely reflects a context-dependent local DNA conformation experienced by the polymerase as it moves into the CGG repeat region.”), Fig. 3 (discussing difference between CGG and AGG polymerase kinetics). While Loomis observed polymerase kinetics on unmodified DNA, the

authors describe their analyses as “establish[ing] a clear foundation for the detection of epigenetic modifications,” which would include detection of modified nucleotides as recited in the claims. EX1009, 126; *see also supra* Sections V.B-C.

86. The combination of kinetic parameters and sequence context was also used for modification detection in related contexts. For example, models for other third-generation sequencing technologies included both kinetic parameters and ordered nucleotide identities for a window of nucleotides in an input data structure for a model (such as a recurrent or convolutional neural network) to detect base modifications. EX1022, Abstract, 8-9, Fig. 1; *see* EX1017, 1-2 (CNNs and RNNs known neural networks for use in genomics). Skilled artisans appreciated that these workflows, while developed for ONT data, were applicable to identification of modified bases in SMRT data. *See* EX1038, 2 (likening the way that both PacBio and ONT detect modifications by measuring kinetics).

87. A person of ordinary skill in the art would have reasonably expected success in achieving the claimed invention as it requires no more than using pulse data generated in the normal course of SMRT sequencing in the manner taught by the Flusberg Publications. Additionally, advances in computing power and deep learning networks that occurred in the ~9 years between Flusberg’s disclosure and the ’794 patent further confirm that a person of ordinary skill in the art would have reasonably expected success. Increased computing resources enabled more routine

and widespread use of data intensive models and processes, including in genomics applications. *E.g.*, EX1022, 8 (describing prediction of methylation based on 7 features provided by ONT sequencing); EX1038, 3-4 (“To improve the performance of the algorithm, we then examined whether a combination of the features might improve the prediction accuracy, finding that the combination of the 3 features (base quality, mismatch and deletion frequency) increased the accuracy of the model (88-91%)”).

88. While I note that the claims are not limited to any particular model, the availability of open-source deep learning networks and the large number of models successfully applied to the analysis of genomic data speak to the relative ease of implementation of such models. The greater availability of computing resources and wider adoption of complex computer models would have further motivated a person of ordinary skill in the art to apply such models to the detection of DNA modifications using the kinetic parameters and sequence information taught by the Flusberg Publications. *See also supra* Section V.D. For example, the prior art at the time expressly embraced neural networks for detecting methylation status including the Keras framework used by Patent Owner. EX1017, 16 (cataloging deep learning resources, including Keras, for genomics applications); EX1035, 3, (“implemented by Keras framework”), 5 (describing “an intelligent and robust method” for methylation detection); EX1022, Fig. 1 (illustrating neural

network for detecting methylation in ONT data); *see also* EX1030, 10 (“The program for the CNN model was implemented on the basis of the Keras deep learning framework (<https://keras.io/>).”).

89. A person of ordinary skill in the art would have reasonably expected success in combining the Flusberg Publications and PacBio-BAM to achieve the claimed invention. The availability of open-source deep learning networks and the large number of models successfully applied to the analysis of genomic data speak to the relative ease of implementation of such models. *See supra* Section V.D. Indeed, the prior art at the time expressly embraced complex models, such as neural networks, in detecting nucleotide modifications. EX1035, 3, (“implemented by Keras framework”), 5 (describing “an intelligent and robust method” for methylation detection); EX1022, 8, Fig. 1 (illustrating neural network for detecting methylation in ONT data); EX1038, 3-4 (discussing accuracy improvements from “inclusion of all the features from the neighbouring positions”). More particularly, the prior art expressly embraced neural networks for detecting methylation status including the Keras framework that was later reported as the framework used by Patent Owner to implement the claimed model. EX1030, 10 (“The program for the CNN model was implemented on the basis of the Keras deep learning framework (<https://keras.io/>).”); *see* EX1017, 16 (cataloging deep learning resources, including Keras, for genomics applications); EX1035, 3 (“implemented by Keras

framework”), 5 (describing “an intelligent and robust method” for methylation detection); EX1022, Fig. 1 (illustrating neural network for detecting methylation in ONT data).

90. In light of the established success in employing such models in the prior art, as well as the specific disclosure of the Flusberg Publications and PacBio-BAM regarding the use of the properties recited in the claims for detecting nucleotide modifications, a person of ordinary skill in the art would have had a reasonable expectation of success in implementing the claimed method.

1. Preamble

91. The Flusberg Publications disclose or teach the preamble of claim 1, which recites “A method for detecting a modification of a nucleotide in a nucleic acid molecule, the method comprising.”

92. Flusberg-338 teaches “single molecule sequencing methods ... for identification of modified nucleotides within nucleic acid sequences.” EX1007, Abstract. For example, Flusberg-338 teaches using “a classifier for distinguishing methylated from unmethylated cytosines in a nucleic acid template” based on a set of pulses’ features in a “real-time sequencing-by-synthesis trace.” EX1007, 66:46-67:4. Flusberg likewise teaches “direct detection of modified nucleotides in the DNA template” for various types of modification. EX1006, Abstract; *see id.*, 463-64. The Flusberg Publications thus teach detecting a modification of a nucleotide

in a nucleic acid molecule (such as a DNA molecule); accordingly, the Flusberg Publications teach the preamble of claim 1.

'794 patent	Prior Art
[Preamble] A method for detecting a modification of a nucleotide in a nucleic acid molecule, the method comprising:	“We describe the direct detection of DNA methylation, without bisulfite conversion, through single-molecule, real-time (SMRT) sequencing. In SMRT sequencing, DNA polymerases catalyze the incorporation of fluorescently labeled nucleotides into complementary nucleic acid strands. The arrival times and durations of the resulting fluorescence pulses yield information about polymerase kinetics and allow direct detection of modified nucleotides in the DNA template, including n6-methyladenine, 5-methylcytosine and 5-hydroxymethylcytosine.” EX1006, Abstract. “Methods, compositions, and systems are provided for characterization of modified nucleic acids. In certain preferred embodiments, single molecule sequencing methods are provided for identification of modified nucleotides within nucleic acid sequences.” EX1007, Abstract. “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.” EX1007, 66:46-67:4. <i>See also</i> EX1006, 463-64.

93. Accordingly, the Flusberg Publications teach the preamble of claim 1.

2. Element 1.1

94. Element 1.1 of claim 1 recites as follows:

(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;

The Flusberg Publications disclose or teach this element.

95. Both Flusberg and Flusberg-338 teach receiving the data recited in element 1.1 from optical signals generated from the incorporation of nucleotides in PacBio single-molecule real-time (SMRT) sequencing. Flusberg illustrates in Fig. 1 (reproduced below) a trace from SMRT sequencing of a nucleic acid molecule that shows all four of the properties recited in element 1.1.

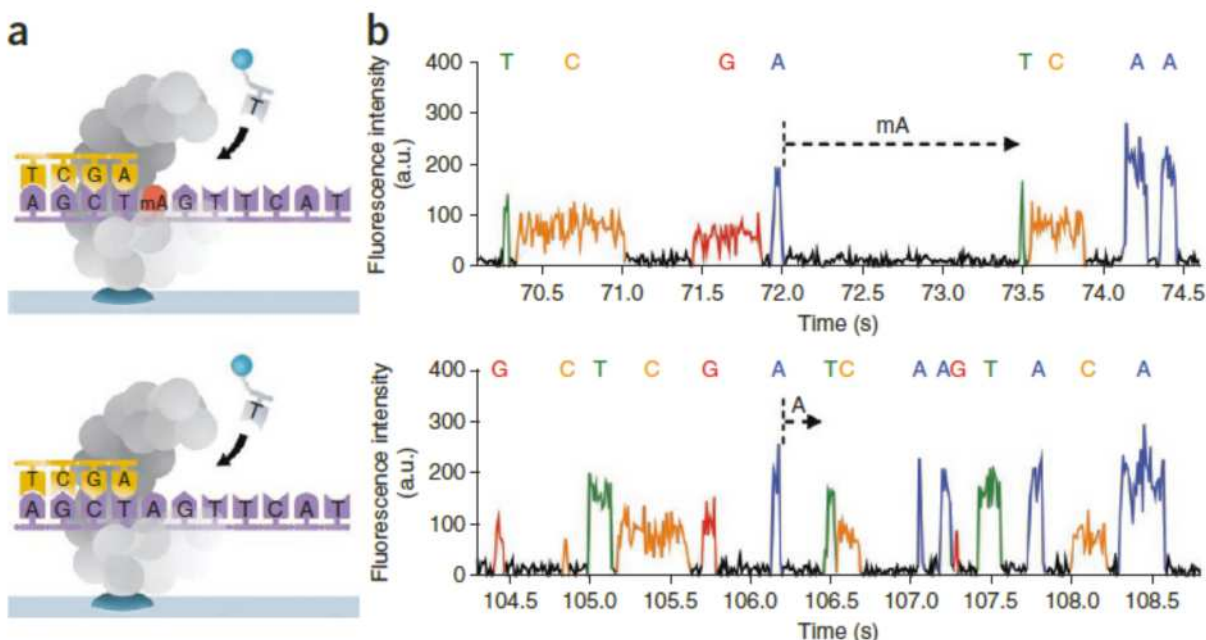


Figure 1 | Principle and corresponding example of detecting DNA methylation during SMRT sequencing. (a) Schematics of polymerase synthesis of DNA strands containing a methylated (top) or unmethylated (bottom) adenine. (b) Typical SMRT sequencing fluorescence traces for samples in a. Letters above the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

EX1006, Fig. 1. In SMRT sequencing, as each nucleotide is incorporated into a nascent strand, the incorporated nucleotide emits a pulse of light from a fluorescent dye tag attached to the nucleotide. EX1006, 462, 466. This pulse of light has a different color for each nucleotide (A, C, T, or G), so the color of emitted light identifies the incorporated nucleotide. EX1006, 462. Because nucleotides are incorporated using a sample strand as a template, the sequence of the template may be inferred from the sequence of pulses emitted from incorporated nucleotides by complementarity. *See id.* Fig. 1 shows the different colors of each pulse, along with the corresponding nucleic acid identity under the principle that the pulse's a

pulse's "color identifies that nucleotide." *Id.* Further, the order of the pulses provides the order of nucleotides in sequence; thus, both the identity of each nucleotide and the position of the nucleotide within the sample nucleic acid molecule are received, as depicted in Fig. 1 (showing colored nucleotide identities in order). *Id.*; *see id.*, Fig. 1.

96. The figure also shows the two remaining properties recited in element 1.1: pulse width (PW) and interpulse duration (IPD). As Flusberg explains, a given nucleotide's "pulse ends when the fluorophore, linked to the nucleotide's terminal phosphate, is cleaved by the polymerase before translocation to the next base in the DNA template." EX1006, 462. The PW and IPD both depend on polymerase kinetics: PW is the time from nucleotide binding to cleavage of the fluorophore by the polymerase, and IPD is the time between two nucleotides' pulses, reflecting the amount of time it takes for the polymerase to shift from one nucleotide incorporation to the next. *See id.* IPD of one of the nucleotides is illustrated in Fig. 1 of Flusberg with a dashed line for each trace, with the remaining IPDs being reflected, though not separately highlighted, by the corresponding spaces between all other pulses. PW is similarly shown (though not highlighted) by the width of each colored pulse. *See* EX1006, Fig. 1. Flusberg thus teaches receiving data acquired by measuring pulses in an optical signal corresponding to nucleotides sequenced in a sample nucleic acid, including obtaining values for properties

including nucleotide identity, position within the sample nucleic acid molecule, pulse width, and interpulse duration.

97. Flusberg-338 provides a similar description of SMRT sequencing and the data received from such sequencing. As Flusberg-338 explains, for each nucleotide in sequence, SMRT sequencing “distinctively identifies the base (e.g., A, G, T, or C)” incorporated into a nascent strand, permitting “the practitioner to identify the base incorporated and, by complementarity, the sequence of the template DNA molecule.” EX1007, 58:32-44. Data is produced by the sequencing reaction, allowing analysis of the kinetics based upon features of the pulses. *See* EX1007, 61:7-28 (describing “Analysis of the data generated by the methods described herein” by “software and/or statistical algorithms,” including “basecalls” and sensitivity to “the presence of modified or damaged nucleotides”), 61:46-54 (detection of pauses during nucleotide incorporations), 66:46-67:4 (tabulation of “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” for methylation detection).

98. Flusberg-338 specifically identifies IPD and pulse width as being values measured from optical sequencing traces that are affected by the presence of nucleotide modifications, both at a modified nucleotide and at nearby nucleotides. EX1007, 22:49-54. Flusberg-338 also presents a simple figure showing how

nucleotide identities, context (identities located together in an ordered sequence), pulse width (PW) and interpulse duration (IPD) may be measured in a pulse trace. *See* EX1007, Fig. 2. Flusberg-338 teaches that pulse features including PW and IPD may be measured for each nucleotide in a local context, and the set of measured values may be used to identify modifications using a model to classify nucleotides as modified or unmodified. EX1007, 66:46-67:4.

99. Flusberg-338 identifies a set of measurable features to be tabulated for the purposes of discriminating between modified and unmodified nucleotides as including “pulse width, pulse duration,” and other features. *Id.* Flusberg-338’s use of the term “pulse duration” would be understood by a person of ordinary skill in the art to refer to IPD—that is, duration *between* pulses, not duration *of* a pulse (which would be a pulse width, redundant with the pulse width listed immediately before “pulse duration”). This interpretation is confirmed throughout the specification of Flusberg-338, as IPD is consistently identified as an important pulse feature affected by and predictive of modifications. *See, e.g.,* EX1007, 67:64-68:5 (“[I]t will be clear to one of ordinary skill that the approach can be extended to pulse metrics other than IPD, such as pulse width....”), 67:25-27 (“In certain embodiments, the set of feature inputs is restricted to those that are most correlated with the Me⁺ and Me⁻ status of a position.”), 35:14-16 (“method to distinguish between C, 5-MeC, and 5-hmC by monitoring deviations from normal polymerase

kinetics, including IPD and pulse width”). By contrast, “pulse duration” is used only once, in the discussion at the bottom of column 66, making it clear that the “pulse duration” being referred to is the *interpulse* duration.

'794 patent	Prior Art
[1.1] (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;	“Incorporation of a nucleotide is detected as a pulse of fluorescence whose color identifies that nucleotide. The pulse ends when the fluorophore, linked to the nucleotide’s terminal phosphate, is cleaved by the polymerase before translocation to the next base in the DNA template. ... Fluorescence pulses in SMRT sequencing are characterized not only by their emission spectra but also by their duration and by the interval between successive pulses. These metrics, defined here as pulse width and interpulse duration (IPD), add valuable information about DNA polymerase kinetics.” EX1006, 462 (citation omitted). “[A] template-directed DNA sequencing reaction in the presence of a polymerase and a set of nucleotides, each of which bears a label that is optically detectable and that distinctively identifies the base (e.g., A, G, T, or C). Detection of an optical signal upon interaction with the polymerase and incorporation into the nascent strand allows the practitioner to identify the base incorporated and, by complementarity, the sequence of the template DNA molecule.” EX1007, 58:32-44. “Analysis of the data generated by the methods described herein is generally performed using software and/or statistical algorithms that perform various data conversions, e.g., conversion of signal emissions into basecalls, conversion of basecalls into consensus sequences for a nucleic acid template, and conversion of various aspects of the basecalls and/or

	consensus sequence to derive a reliability metric for the resulting values.” EX1007, 61:7-28. “For example, the presence of modified bases in a template nucleic acid has been shown to change the width of a pulse and/or the interpulse duration (IPD), either at the modified base or at one or more positions proximal to it.” EX1007, 22:49-54. <i>See also</i> EX1006, 463, 466, Figs. 1, 3; EX1007, 35:14-16, 61:46-54, 66:46-67:4, 67:25-27, 67:64-68:5, Fig. 2.
--	---

100. Accordingly, the Flusberg Publications teach element 1.1.

3. Element 1.2

101. Element 1.2 of claim 1 recites as follows:

(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:

an identity of the nucleotide,

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

the width of the pulse corresponding to the nucleotide, and

the interpulse duration;

The Flusberg Publications in view of PacBio-BAM disclose or teach this element.

102. To begin with, regarding “creating an input data structure ... comprising a window of the nucleotides sequenced in the sample nucleic acid molecule,” the specification of the ’794 patent teaches that a window may be symmetric or asymmetric around the target nucleotide, with unlimited variation in

size. EX1001, 27:39-51 (“A stretch of DNA surrounding a base under interrogation could be X-nt upstream and Y-nt downstream of that base, namely the ‘measurement window.’ X could include, but is not limited to, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 100, 150, 200, 300, 400, 500, 1000, 2000, 4000, 5000, and 10000; Y could include, but is not limited to, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 100, 150, 200, 300, 400, 500, 1000, 2000, 4000, 5000, and 10000.”), 33:12-48. Furthermore, the measurement window could be the set of all nucleotides in a data structure storing a template molecule: “measurement windows may be the entire template molecule.” EX1001, 34:46-47. Thus, the recited “window” may encompass any set of nucleotides in a data structure—either all of the nucleotides in the data structure or any subset of the nucleotides.

103. Flusberg-338 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. EX1007, 66:46-67:4. Flusberg-388 teaches performing “tabulation” of “local pulse features” over a “local context”

around a “pulse of interest,” forming a “feature vector” $\vec{f}$ that is written in the form

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

104. A person of ordinary skill in the art would understand from Flusberg-338 that such types of data structure would be input data structures, as the feature vector $\vec{f}$ is an input for a model that distinguishes methylated and unmethylated nucleotides. *See* EX1007, 66:46-67:36. In particular, Flusberg teaches using the $\vec{f}$ vector as an input to produce “a generalized signal for determining methylation status on the trace,” in the form

$$w_i = \log \left(\frac{p(\vec{f} | Me+)p(Me+)}{p(\vec{f} | Me-)p(Me-)} \right)$$

Id. The use of the vector $\vec{f}$ in the numerator and denominator of the fraction shows that the data structure $\vec{f}$ is an input to Flusberg-338’s classifier, as it is used to produce an output signal for determining methylation status from the pulse trace for which it represents a portion corresponding to a window of nucleotides. The specific form of the equation given for Flusberg-338’s model indicates that it is a Bayesian odds-ratio, in logarithmic form, meaning that by using the feature vector $\vec{f}$ as an input, in conjunction four derived values (the prior probabilities of methylation and non-methylation, and the prior probabilities of seeing a feature vector $\vec{f}$ for a target nucleotide with or without methylation), a value is produced that represents the logarithm of the probability ratio of the actual strand in question

being methylated or non-methylated, given the measured vector $\vec{f}$. If the probability of it being methylated exceeds the probability of it being not, the odds ratio will be greater than 1, and thus the log of odds ratio will be a positive value. Conversely, if the probability of the strand being methylated is below that of it not being methylated, the odds ratio will be smaller than one, and thus the log of it will be a negative value. An odds ratio of 1, and thus a log of odds ratio of 0, would indicate that the two probabilities are equal.

105. With regard to the specific number of nucleotides in a window, Flusberg-338 suggests a tabulation of the features of each pulse over a “local context extending 3 pulses to the left and right of the pulse of interest,” but notes that “context size is flexible and, in certain embodiments, can be application-specific.” EX1007, 66:46-67:4. The feature vector $\vec{f}$ is precisely a “data structure comprising a window of the nucleotides sequenced in the sample nucleic acid molecule.” A POSA would understand that the vector corresponds to a specified subset of nucleotides within a larger sequence (e.g., specified portion of a BAM file containing many sequencing reads for each read a sequence of nucleotides, and for each nucleotide in a read the associated kinetic parameters). A POSA would also understand that, as convenient, such a vector could be stored in a data structure (such as a BAM file) either with or without features for other nucleotides from a larger sequence.

106. In Flusberg-338’s example, the window would encompass the set of nucleotides in the local context (e.g., from nucleotide $i-3$ to $i+3$). The target nucleotide would be the one corresponding to the pulse labeled i . The other nucleotides in the window are identified by their distance from the target nucleotide; for example, pulse $i-3$ is the nucleotide 3 bases in one direction from the target, while pulse $i+3$ is the nucleotide 3 bases in the other direction. The ordered list of which the vector is formed is a data structure, since it is a structure containing each relevant pulse feature for each nucleotide (these features being indicated in vector $\vec{f}$ with subscript $1, 2, \dots, k$). Flusberg-338 also refers to it as the product of a tabulation, corresponding to a data structure in the form of a table. EX1007, 66:46-67:4; *see also id.*, 64:29-65:20 (set of “observed pulse data (both basecalls and other pulse features)” used to detect pauses, which can be associated with the “occurrence of modified nucleotide bases”). A person of ordinary skill in the art would understand that the same vector of data could be stored in various interchangeable ways in a computer file, including as a 1-dimensional array, a two-dimensional array (e.g., a $k \times w$ matrix, where w is the window size (such as 7) and k is the number of features for each pulse), a set of k arrays (one for each pulse feature, with w pulse elements per array, as taught by PacBio-BAM), a set of w arrays (one for each nucleotide, with k different pulse features per array), etc. *See id.* Indeed, a person of ordinary skill in the art find it natural to create such data

structures, including the structure disclosed by PacBio-BAM, in order to use it as an input for a classifier. Likewise, a POSA would know that such a data structure could be formed to contain exclusively the features for a particular window of nucleotides or that the data structure could encompass a larger sequence of nucleotides, with the window being a specified subset within the data structure's stored nucleotides, using programmatic techniques such as a sliding window to read from a subset of a sequence. *See, e.g.*, EX1038, 5-7.

107. Flusberg similarly teaches that “the kinetic signatures of methylation will be sequence context-dependent,” proposing that “detection of methylation may also be possible by tabulating expected kinetics over a suitable number of contexts or by taking advantage of heuristics that embody the observed trends in SMRT sequencing kinetics.” EX1006, 465. By proposing tabulating kinetics over different contexts to detect methylation, Flusberg is suggesting that the specific sequence of a window of nucleotides (the “context”) be an input to a model for detecting methylation, because the context affects kinetic signatures including IPD and pulse width. 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 nucleotide sequence. *See* EX1006, 465 (showing same nucleotide sequences used for Figs. 1-4, with only modification status varied). Thus, the

model learned IPD and PW principal components specific to that sequence, but Flusberg recognizes that a different sequence (with different nucleotide identities and positions relative to the nucleotide of interest) would have a different set of principal components. By proposing tabulating kinetics over different contexts, Flusberg is proposing developing a classifier that uses context (nucleotide identity and relative positions in a window) as an input variable, along with IPD and PW, so as to classify methylation status based on a combination of all of these variables. Accordingly, Flusberg confirms that a data structure for detecting methylation should contain each of properties recited in element 1.2 for each nucleotide within a window.

108. In light of the Flusberg Publications' teachings, a POSA seeking to detect modifications would have created a data structure to store relevant features for a window of nucleotides. This structure could contain only features for the window, or it could store many nucleotides' features with only a specified subset being the window.

109. Regarding the form of data structure used for the tabulated vectors of the Flusberg Publications, PacBio-BAM teaches creating an input data structure (a BAM file), which, as I discuss below, contains each of the properties recited in element 1.2 for a window of nucleotides.

110. **“an identity of the nucleotide”**: PacBio-BAM teaches creating a data structure (a BAM file) containing each of the recited properties for a window of nucleotides. Regarding nucleotide identities and positions, PacBio-BAM states that “PacBio-produced BAM files are fully compatible with the BAM specification.” EX1011, 1. This sentence includes a link to download the SAM/BAM specification (EX1012) at <http://samtools.github.io/hts-specs/SAMv1.pdf>. The BAM specification referred to by PacBio-BAM teaches that in each BAM file, the nucleotide “[s]equence is encoded in 4-bit values, with adjacent bases packed into the same byte starting with the highest 4 bits.” EX1012, 15. For the particular encoding provided, it explains that the “case-insensitive base codes ‘=ACMGRSVTWYHKDBN’ are mapped to [0; 15] respectively with all other characters mapping to ‘N’ (value 15).” *Id.* This means that the characters in the string “=ACMGRSVTWYHKDBN” are encoded (with regard to upper- or lowercase) as a number from 0 to 15, with 0 being “=”, 1 being “A”, etc., up to “N” being 15. The “4-bit” value refers to an encoding with four binary digits, so 0000 is zero, 0001 is one, 0010 is two, and 1111 is fifteen. In this encoding, an adenine nucleotide (A) would be 0001 (one in decimal), a cytosine nucleotide (C) would be 0010 (two in decimal), a guanine nucleotide (G) would be 0100 (four in decimal), and a thymine nucleotide (T) would be 1000 (eight in decimal). In combination, then, a nucleotide sequence would be encoded as an array of these

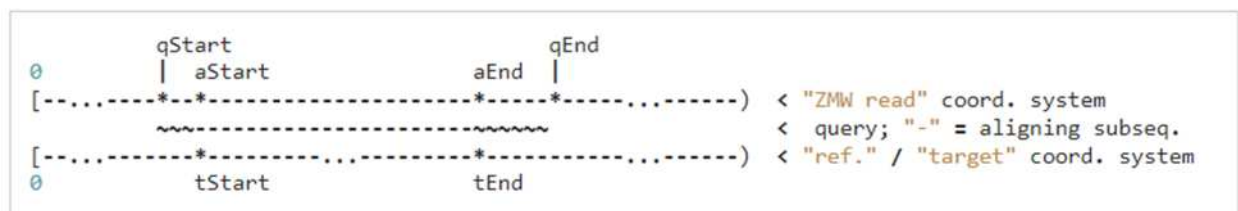
four-bit values. For example, the sequence GATC would be encoded as the bit sequence 0100000110000010, forming an array of four 4-bit numbers, i.e., (0100,0001,1000,0010), with the nucleotide identified by which of a given set of four bits was a one, and the order of the 4-bit integers indicating the order of bases in the nucleotide sequence.

111. Including the identity of each nucleotide in the data structure, as taught by PacBio-BAM, is consistent with the Flusberg Publications' teachings. Flusberg-338 teaches that sequence context is a relevant "feature," noting that "the local sequence context of the template strand can also influence and inhibit the activity of the polymerase along the template." EX1007, 62:28-39; *see id.*, 64:53-55 ("It can also better integrate knowledge of the inter-pulse spacing and other information, such as sequence context, into a broad model."), 66:4-45 (probability of methylation related to "neighboring sequence context" and features such as IPD), 5:33-46 (teaching a "combination of sequence context and methylation status affected interpulse duration" when detecting methylation); *see also id.*, 7:4-17 (pause indicative of modification), 64:29-65:20 (prediction based on "observed pulse data (both basecalls and other pulse features)"). Indeed, Flusberg-338 expressly refers to basecalls as being a type of pulse feature used for a model making predictions indicating nucleotide modifications. *See id.*, 64:29-65:20

(prediction based on “observed pulse data (both basecalls and other pulse features)”), 7:4-17 (pause indicative of modification).

112. **“a position of the nucleotide with respect to a target position within the window”**: PacBio-BAM includes the relative positions of each nucleotide, indicating where each nucleotide is in sequence, and thus relative to a nucleotide of interest in the window. As I discussed above, PacBio-BAM stores nucleotide identities in an array, which provides the position of each nucleotide in sequence (an array being a data structure that encodes both the members of a set of data objects and the relative position of each object by storing them in order).

113. PacBio-BAM indicates that the nucleotide sequence encoded in a BAM file may either be unaligned or aligned to a reference genome. EX1011, 1-2. The BAM file includes a position of the nucleotide with respect to a target position within a window by aligning start and end positions within a sequence read (qStart, qEnd) to a target sequence (tStart, tEnd), as shown in the figure reproduced below:



EX1011, 1-2. As PacBio-BAM explains, “sequence read presented to an aligner is termed a *query*,” which will typically “be a subsequence of an entire PacBio ZMW read.” *Id.* The query can be soft-clipped, meaning that although “entirety of the

query is stored in the aligned BAM,” a field indicates that only a subset of bases are included in the alignment. *Id.* As the figure above depicts, the BAM file includes a query running from qStart to qEnd, and a window from aStart to aEnd is defined within the query, with the window being those bases aligning to a target sequence running from tStart to tEnd.

114. Identifying the position of each nucleotide relative to a target nucleotide is also consistent with the Flusberg Publications. As I discussed above, Flusberg-338’s feature vector labels each nucleotide’s set of features with a superscript indicating its relative position, with superscript i representing the target nucleotide’s pulse, and the other nucleotides being labeled from $i-3$ to $i+3$. *See* EX1007, 66:46-67:4. These labels are direct indications of each nucleotide’s position with respect to the target position (nucleotide i) within the window.

115. Flusberg similarly discloses tracking each nucleotide’s relative position. Flusberg proposes “that methylated bases in a DNA template might be detected directly on the principle that their presence affects polymerase kinetics during SMRT sequencing.” EX1006, 462. It teaches that pulse width and IPD “add valuable information about DNA polymerase kinetics,” and it shows that a combination of IPD and pulse width allows differentiation between unmodified cytosine, 5-methylcytosine, and 5-hydroxymethylcytosine. *See* EX1006, Fig. 3 (showing that the first two principal components determined from measured data

differentiate between these three methylation states); *see also infra* Section XI.A.4 (discussing principal component analysis in more detail). To differentiate between these different types of unmodified and modified cytosines, Flusberg uses principal component analysis to find a group of weights for the respective pulse widths and IPDs of each of a set of nucleotides in a window. *See* EX1006, 462; *see also* EX1001, claim 11 (identifying “principal component analysis” as a type of model). As I explain below in more detail for element 1.3, Flusberg’s supplementary materials indicate, similar to Flusberg-338, that each weight is identified with a particular nucleotide based on its position relative to the nucleotide of interest, identified as nucleotide 0. *See* EX1006, 478; *infra* Section XI.A.4.

116. **“the width of the pulse corresponding to the nucleotide, and the interpulse duration” (i.e., PW and IPD):** In addition to encoding the identity and relative locations of each nucleotide, PacBio-BAM discloses encoding the pulse width and IPD of each nucleotide. Pulse width and IPD values for each nucleotide are encoded in respective arrays of integers, either in a lossless format as an array of 16-bit integers or in a compressed format as an array of 8-bit integers. EX1011, 8. PacBio-BAM further indicates that these arrays are to be used for downstream analysis, noting that either format of encoding for IPD and pulse width (lossless or compressed) is appropriate for “important production use cases,” and specifically for “base modification detection.” EX1011, 8.

117. Each of the array of nucleotide identities, array of the pulse widths and the array of IPDs is a data structure. An array is one of the most fundamental forms of data structure in computing, taking the form of a collection of items stored at contiguous memory locations. For an integer array, the structure is formed by encoding each integer using a specified number of bits (4 bits for nucleotides, 8 or 16 bits for PW and IPD values), and then placing the binary-encoded integers into sequential memory locations, thereby forming an array data structure. Moreover, a file is a type of data structure, so a file like the BAM file disclosed by PacBio-BAM that contains multiple arrays is itself a data structure. Thus, as PacBio-BAM discloses a file containing arrays of PW, IPD, and nucleotide identity values in sequential order for a sequence of nucleotides (i.e., a window of nucleotides), PacBio-BAM is a data structure containing all properties recited in element 1.2: for each nucleotide in a window of nucleotides, it contains a PW value, an IPD value, and the nucleotide's identity, as well as indicating the position of each nucleotide relative to each other nucleotide by encoding the nucleotide identities (and other properties) in an ordered array reflecting the nucleotides' positions in sequence. Indeed, as I noted above in Section VIII, Take2's infringement contentions indicate that PacBio BAM files contain each property recited in element 1.2, relying on the 2022 PacBio-BAM specification,

which has substantially the same disclosure as PacBio-BAM with respect to these properties.

118. The inclusion of PW and IPD as features in an input data structure is further supported by the Flusberg Publications. Flusberg-338 teaches that the feature vector contains features measured for each pulse, and it specifically teaches that both PW and IPD may be included. *See* EX1007, 66:46-67:4 (“For example, a set of measurable features might be pulse width, [inter]pulse duration, ...”). Flusberg-338 repeatedly confirms that PW and IPD are features that are affected by the presence of nucleotide modifications, both at the modification site and for nearby nucleotides. *See* EX1007, 9:41-46; 22:49-54; Figs. 12-13, 15-19 (showing PW and IPD ratios as a function of nucleotide position).

119. Flusberg agrees. As I discussed above, Flusberg teaches using PW and IPD to differentiate between unmodified cytosine, 5-methylcytosine, and 5-hydroxymethylcytosine by inputting them into a principal component analysis model for each nucleotide positioned around a target nucleotide in a window of nucleotides. *See* EX1006, 462, Fig. 3.

'794 patent	Prior Art
[1.2] (b) creating an input data structure, the input data structure comprising a window of the nucleotides	“PacBio-produced BAM files are fully compatible with the BAM specification.” EX1011, 1. “Sequence is encoded in 4-bit values, with adjacent bases packed into the same byte starting with the highest 4 bits first. For a sequence with an odd

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;	numbered length the bottom 4 bits of the last byte are undefined, but we recommend writing these as zero. The case-insensitive base codes ‘=ACMGRSVTWYHKDBN’ are mapped to [0; 15] respectively with all other characters mapping to ‘N’ (value 15)”. EX1012, 15. “Interpulse duration (IPD) and pulsewidth are measured in frames; natively they are recorded as a uint16 per pulse/base event. They may be encoded in BAM read tags in one of two fashions: • losslessly as an array of uint16; necessary for PacBio-internal applications but entails greater disk space usage. • lossy 8-bit compression stored as a uint8 array, following the codec specified below (“codec V1”). Provides a substantial disk-space savings without affecting important production use cases (base modification detection).” EX1011, 8. <i>See also</i> EX1011, 1-8; EX1006, 462, 465, Fig. 3, Supplementary Table 1; EX1007, 5:33-46, 7:4-17, 9:41-46; 22:49-54, 62:28-39, 64:53-55, 66:4-45, 66:46-67:4, Figs. 12-13, 15-19.
--	--

120. Accordingly, the Flusberg Publications in view of PacBio-BAM teach element 1.2.

4. Element 1.3

121. Element 1.3 of claim 1 recites “(c) inputting the input data structure into a model, the model trained by:.” The Flusberg Publications in view of PacBio-BAM disclose or teach this element.

122. For example, Flusberg-338 teaches “a classifier for distinguishing methylated from unmethylated cytosines in a nucleic acid template.” EX1007, 66:46-67:36. Flusberg-338 further discloses that “a generalized signal for determining methylation status on the trace” is derived by measuring the “prior probabilities of methylated or un-methylated positions” and the “observed data likelihoods” of observing a feature vector given a methylated or unmethylated position. *Id.* As I explained above in my discussion of element 1.2, Flusberg-338 is describing a model in which the feature vector $\vec{f}$ (i.e., the data structure discussed for element 1.2) is an input and the output is a Bayesian odds ratio—the ratio of the probability of the target nucleotide being methylated to the probability of the target nucleotide being unmethylated—expressed as a logarithm.

123. Further, Flusberg-338 teaches the use of many types of model beyond the one I discussed above for element 1.2. For example, Flusberg-338 discloses that its “general classifier ... can be implemented with many standard statistical classification algorithms, i.e. linear discriminant analysis, multi-dimensional regression, kernel methods, classification and regression trees, neural networks, and support vector machines.” In the context of the ’794 patent claims, each of Flusberg’s various classifiers is a model. The ’794 patent identifies various types of classifiers as models, including a “Bayes classifier,” a “neural network” and a “support vector machine.” EX1001, 89:25-51. The first of these examples would

correspond to the classifier I discussed above for element 1.2, and the latter two examples are expressly identified by Flusberg-338 as alternatives for its classifier. *See, e.g.*, EX1007, 65:13-16 (“Bayesian likelihood model for identifying modified bases”), 66:46-67:36, 67:67-68:5.

124. A person of ordinary skill in the art would understand that a model like Flusberg-338’s classifier would receive a data structure as taught by Flusberg-338 as an input and would be trained via data structures containing the same elements, for which the methylation status is known. For example, the probabilities used for the classifier of Flusberg-338 that I discussed above for element 1.2, such as the probabilities of seeing the vector $\vec{f}$ given that the target nucleotide was methylated or unmethylated, would be determined by training the model, i.e., by presenting examples of such vectors to the model to allow it to derive the appropriate probabilities statistically. Flusberg-338 identifies two types of training techniques for such a model—“a kernel density method or simple binning and tabulation of the features”—and further teaches that “[v]arious standard classification algorithm development techniques known to those of ordinary skill in the art may be applied to refine this approach, both to reduce training set bias and to improve sensitivity.” EX1007, 67:6-36. Flusberg-338 thus teaches training its model by inputting data structures for methylated and non-methylated nucleotide sequences that match the type of data structure used for classification;

indeed, a person of ordinary skill in the art would have understood that it is by inputting such data structures (labeled appropriately as methylated or non-methylated) that a classifier can learn to distinguish methylated and non-methylated sequences from pulse features like PW, IPD, and ordered nucleotide identities. Likewise, a person of ordinary skill in the art would have known that the other examples of classifiers identified by Flusberg 338 would also be trained in the same way; for example, neural networks have long been trained by subjecting them to labeled training data in structures matching the data structures they are to operate on in implementation, and the other examples from Flusberg-338 are trained similarly.

125. Flusberg similarly teaches inputting an input data structure into a trained model. For example, it teaches using a principal component analysis model that takes a data structure including IPD and pulse width signals over a window of nucleotides as input and outputs a signal discriminating between different modification states (such as unmodified, methylated, and hydroxymethylated) at a given nucleotide. EX1006, 463. This model includes weights for each nucleotide and pulse feature derived by fitting to training data. *Id.* The weights are shown in Flusberg's Supplementary Table 1, which includes multipliers for the IPD and PW of each of 9 nucleotides (numbered -2 through 6), identified by their positions relative to a target nucleotide (nucleotide 0).

126. These multipliers are weight values derived via principal component analysis, in which a set of training data is fit to a model to derive a series of “principal components.” Explained simply, the first “principal component” is a line of best fit to the data, the second principal component is a line of best fit to the data subject to the requirement that it be orthogonal to the first principal component, and each subsequent principal component is similarly a line of best fit that must be orthogonal to all prior components. Thus, a set of N principal component is a model of data, with each added component explaining a smaller and smaller amount of the data’s variance. *See also* EX1001, claim 11 (model includes principal component analysis). For Flusberg, inputting a set of measured pulse features into a set of N principal components will produce an N-dimensional value differentiating methylation states; for example, Fig. 3 of Flusberg shows that using just the first two principal components can produce a 2-dimensional value differentiate between unmodified, methylated, and hydroxymethylated versions of a give sequence of nucleotides. *See* EX1006, Fig. 3 (showing projection of data onto first two principal components, producing values that separate between these three modification states in a 2-dimensional graph).

’794 patent	Prior Art
[1.3] (c) inputting the input data structure into	“In certain aspects, the invention provides a general-purpose approach to discriminating between Me ⁺ and Me [–] using features in a real-time sequencing-by-

a model, the model trained 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. The observed data likelihoods $p(\vec{f} Me +)$ and $p(\vec{f} Me -)$ are derived, e.g., by a kernel density method or simple binning and tabulation of the features. Thus, a generalized signal for determining methylation status on the trace has been determined: $w_i = \log \left(\frac{p(\vec{f} Me +)p(Me +)}{p(\vec{f} Me -)p(Me -)} \right)$ where $p(Me+)$ and $p(Me-)$ are the prior probabilities of methylated or un-methylated positions, respectively.” EX1007, 66:46-67:36. <i>See also</i> EX1006, 463, 464-65, Supplementary Table 1; EX1007, 65:13-16, 67:64-68:5, Fig. 20.
---------------------------------------	--

127. Accordingly, the Flusberg Publications teach element 1.3.

5. Element 1.4

128. Element 1.4 of claim 1 recites as follows:

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.

This element describes a training set for the model recited in element 1.3.

According to Element 1.4, the training set must have a plurality of data structures, one for each of a plurality of nucleic acid molecules, and each data structure includes the properties recited in element 1.2 for a window of nucleotides based on their measured pulses. Element 1.4 further requires the nucleotide in a target position to have a known modification state. As I explain below, the Flusberg Publications in view of PacBio-BAM disclose or teach element 1.4 of claim 1.

129. As I addressed for element 1.1, the Flusberg Publications teach measuring pulses to derive properties for a window of nucleotides in a plurality of nucleic acid molecules with known modification states. *See* EX1006, Fig. 1 (showing pulse traces for methylated and unmethylated nucleic acids), 463-65 (describing analysis of properties of methylated and unmethylated nucleic acids),

Fig. 3 (displaying data from pulse features from multiple nucleic acid molecules), Supplementary Note (providing different nucleic acid sequences), Supplementary Table 1 (providing principal components extracted from pulse properties of methylated and unmethylated templates); EX1007, 66:46-67:36 (describing development of classifier on a training set comprising “methylated and unmethylated template data sets” analyzed over a window of features); *see also id.*, 64:29-43, Fig. 20 (providing nucleotide sequences with identified modifications), *supra* Section XI.A.2.

130. 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. As I further discussed for element 1.3, each of these data structures would have a known modification state—modified or unmodified—which would be used as a label for the training data structures. The degree to which the model accurately reproduced the labels of modification state based on

the relevant properties in the training data structures (nucleotide identity, position, PW, and IPD for all nucleotides in the window) would be a measure of how well the model had been trained and would be part of a feedback loop for improving training by modifying parameters of the model. As I described above, both the principal component analysis of Flusberg and the Bayesian statistical analysis of Flusberg-338 are trained in this way. For example, Flusberg-338's method involves training to determine posterior probabilities like the probability of seeing the vector $\vec{f}$ (i.e., the set of values in the data structure) given that the target nucleotide was methylated or unmethylated, which involves learning an association between joint patterns of PW, IPD, and nucleotide sequence values and methylation states. Similarly, other forms of model disclosed by Flusberg, such as neural networks, principal component analysis, etc., involve similar methods of training that would use the same data structures labeled with known modification states.

'794 patent	Prior Art
[1.4] 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	“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 Various standard classification algorithm development techniques known to those of ordinary skill in the art may be applied to refine this approach, both to reduce training set bias and to improve sensitivity. ... In further embodiments, leave-one-out cross-validation

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,	can be valuable in selecting a robust predictive algorithm, e.g., by mitigating overfitting to the observed data that can occur when developing a classifier on a training set.” EX1007, 66:46-67:36. <i>See also</i> EX1006, 463-65, Figs. 1, 3; EX1007, 64:29-43.
--	--

131. Accordingly, the Flusberg Publications teach element 1.4.

6. Element 1.5

132. Element 1.5 of claim 1 recites “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.” The Flusberg Publications in view of PacBio-BAM disclose or teach element 1.5 of claim 1.

133. As I discussed above for elements 1.2-1.4, the Flusberg Publications in view of PacBio-BAM teach creating and receiving the recited first data structures, to be used as samples in a training set for a model to detect nucleotide modifications. A person of ordinary skill in the art would understand that such training samples would be stored for training a model, as taught by the Flusberg

Publications. For example, Flusberg-338 discloses providing “storage for the results of the methods of the invention.” EX1007, 73:35-42; *see also, e.g., id.*, 66:46-67:36, 71:49-72:3, 73:6-25, 73:43-55. A person of ordinary skill in the art would understand that data structures (such as the type of files disclosed by PacBio-BAM) should be stored when forming a training dataset, so that during training, the model can access the data structures to learn to predict modification states from the values of the different properties they contain. This training would be executed by a computer processor, so the data structures would have been stored in memory (such as on one or more hard disks) to allow the processor to read the values stored in each data structure (and typically to do so repeatedly during a continuing training process).

134. As I further explained above, a person of ordinary skill in the art would understand from the Flusberg Publications that the training samples should be labeled indicating the “first state of the nucleotide at the target position.” In other words, the labels for training would indicate whether a modification was present at the target position (and optionally type of modification, such as whether the modification was a methylation or hydroxymethylation). Consistent with this understanding, Flusberg-338 teaches training a classifier based upon separate “methylated and unmethylated template data sets.” EX1007, 66:46-67:36. Both Flusberg Publications include labels indicating the presence of modified bases in

sequences. *See* EX1007, Fig. 20 (labeling modification with asterisks); EX1006, Supplementary Note (labeling modifications with “m” or “hm” for methylation and hydroxymethylation). As I explained in detail above, a person of ordinary skill in the art would understand that for a model to learn to detect modifications from a combination of modified and unmodified samples, the modification state of the nucleotide of interest should be specified so that the ability to correctly detect methylation can be used to judge the model’s accuracy and provide appropriate feedback to learn from. For example, Flusberg-338 teaches determining probabilities of obtaining particular feature vectors given methylation or non-methylation. *See* EX1007, 66:46-67:36. Determining these values requires identifying which samples are methylated, as learning each probability requires learning an association between a vector of properties $\vec{f}$ and a modification state. Likewise, producing the colored principal component graph shown in Fig. 3 of Flusberg requires identifying which data points have which modifications, as each dot is colored *based* on the actual modification state, with the color itself reflecting the label of the modification state. *See* EX1006, Fig. 3 (blue for unmodified, red for methylated, black for hydroxymethylated).

'794 patent	Prior Art
[1.5] storing a plurality of first training samples, each including one of the	“The invention also provides a computer program product comprising a computer-readable medium having a computer-readable program code embodied

first plurality of first data structures and a first label indicating the first state of the nucleotide at the target position, and	therein, the computer readable program code adapted to implement one or more of the methods described herein, and optionally also providing storage for the results of the methods of the invention.” EX1007, 73:35-42. <i>See also</i> EX1006, Fig. 3, Supplementary Note; EX1007, 66:46-67:36, 71:49-72:3, 73:6-25, 73:35-42, Fig. 20.
---	---

135. Accordingly, the Flusberg Publications in view of PacBio-BAM teach element 1.5.

7. Element 1.6

136. Element 1.6 of claim 1 recites as follows:

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

The Flusberg Publications in view of PacBio-BAM disclose or teach this element.

137. Element 1.6 describes a well-known method of training a model, wherein the model’s parameters are optimized so that the output (in this case, the target nucleotide being modified or not modified) best reproduces the labels of the training data. I explained this process above in relation to various types of models disclosed in the Flusberg Publications. *See supra* Section XI.A.4. In the case of the Flusberg Publications, the output of the model would be specifying whether the

nucleotide at the target position has a modification, such as a methylation. For example, Flusberg-338 discloses that by analyzing features of methylated and unmethylated samples, “a generalized signal for determining methylation status on the trace has been determined,” using a vector of features corresponding to the claimed data structure as an input. EX1007, 66:46-67:36. Flusberg-338 further discloses that “standard classification algorithm development techniques known to those of ordinary skill in the art may be applied to refine this approach,” suggesting techniques including “linear discriminant analysis, multi-dimensional regression, kernel methods, classification and regression trees, neural networks, and support vector machines.” EX1007, 66:46-67:36, 67:64-68:5.

138. These well-known machine learning approaches train models by optimizing the models’ parameters based on whether their output matches the labels of the training data. Taking a neural network as an example, popular neural networks include recurrent neural networks (RNNs) and convolutional neural networks (CNNs). *E.g.*, EX1022, 1-2, 8-9 (using RNN to detect methylation); EX1038 (using CNN to estimate probability of methylation); EX1028, 787 (describing RNN used in machine learning), 791 (citing article on “Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures”); EX1017, 1-2 (discussing CNNs and RNNs), 15 (“Most approaches use CNN or RNN....”). A neural network is formed of a sequence of

“layers” of neurons, with the neurons of each layer connected to neurons of the next layer. EX1017, 1-2; EX1035, 3-4. Often, each neuron in a given layer will connect to each neuron in the next layer, forming a “fully connected layer.” EX1035, 3-4. Each of these connections has a “weight” representing the strength of the connection. EX1017, 1-2. When information passes from one layer to the next, it is influenced by these weights. Indeed, the input to a neuron is the combination of outputs from the previous layer's neurons, each multiplied by its respective weight. *See id.* Thus, starting from the input layer (e.g., an input data structure), the data flows through the network, with each layer processing it, making use of these weights, and refining the information until a final prediction is made in the output layer (e.g., modified or unmodified target nucleotide). A key difference between a CNN and an RNN is that an RNN is designed to act on a sequence of input data structures, with the results of the analysis of prior data structures being used as an input to a given data structure later in the sequence. *See id.* By contrast, a CNN produces its output based only on a given input data structure and the weights determined from training, not on the immediately previous input data structure(s) that the model has analyzed. *See id.*

139. The connection weights of a neural network act as tunable parameters, allowing the network to learn and adjust its predictions based on the training data. Specifically, those predictions are compared to a label for that training element to

score the prediction, and the weights are varied to optimize the model's score for predicting all of the elements in the training set. *See* EX1017, 1-2. In the context of the Flusberg Publications, the model's parameters would be varied to optimize the ability to detect the labeled nucleotide modification states of each input data structure in a training set, based on the contents of those data structures. *See also supra* Sections V.C, XI.A.4 (discussing other examples of models and their training).

'794 patent	Prior Art
[1.6] 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	“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. ... The observed data likelihoods $p(\vec{f} Me +)$ and $p(\vec{f} Me -)$ are derived, e.g., by a kernel density method or simple binning and tabulation of the features. Thus, a generalized signal for determining methylation status on the trace has been determined: $w_i = \log \left(\frac{p(\vec{f} Me +)p(Me +)}{p(\vec{f} Me -)p(Me -)} \right)$ where p(Me⁺) and p(Me[−]) are the prior probabilities of methylated or un-methylated positions, respectively. Various standard classification algorithm development techniques known to those of ordinary skill in the art may be applied to refine this approach, both to reduce training set bias and to improve sensitivity. Such techniques include but are not

	limited to cross-validation, boosting, and bootstrap aggregating (bagging). In certain embodiments, the set of feature inputs is restricted to those that are most correlated with the Me⁺ and Me[−] status of a position. In certain embodiments, the major component in a principal components analysis can serve as a better weighted combination of the most important features. In further embodiments, leave-one-out cross-validation can be valuable in selecting a robust predictive algorithm, e.g., by mitigating overfitting to the observed data that can occur when developing a classifier on a training set. Further, in some embodiments a boosting approach (training of a hierarchy of classifiers on the progressively more difficult regions of feature space) is applied to improve sensitivity.” EX1007, 66:46-67:36. “In addition, the general classifier approach suggested in steps 2+3 can be implemented with many standard statistical classification algorithms, i.e. linear discriminant analysis, multi-dimensional regression, kernel methods, classification and regression trees, neural networks, and support vector machines.” EX1007, 67:64-68:5. <i>See also</i> EX1006, 465, 466-67, Table 1.
--	---

140. Accordingly, the Flusberg Publications in view of PacBio-BAM teach element 1.6.

8. Element 1.7

141. Element 1.7 recites “(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.” The Flusberg Publications in view of PacBio-BAM disclose or teach element 1.7 of claim 1.

142. As I discussed above for elements 1.2 and 1.3, Flusberg Publications in view of PacBio-BAM teach inputting the recited input data structure having a window with a nucleotide at a target position. *Supra* Sections XI.A.3-4. And as I discussed for element 1.6, the Flusberg Publications further teach optimizing a model to detect methylation from an input data structure. *See, e.g.*, EX1007, 66:46-67:36; *supra* Section XI.A.7. Element 1.7 corresponds to using the trained model for its intended purpose: to determine whether a nucleotide at a target position is modified. The models taught by the Flusberg Publications take a data structure as an input and produce an output that is a determination of whether the modification is present on the nucleotide at the target position in the window of the input data structure, thus meeting element 1.7.

'794 patent	Prior Art
[1.7] (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.	“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. The observed data likelihoods $p(\vec{f} Me +)$ and $p(\vec{f} Me -)$ are derived, e.g., by a kernel density method or simple binning and tabulation of the features. Thus, a generalized signal for determining methylation status on the trace has been determined: $w_i = \log \left(\frac{p(\vec{f} Me +)p(Me +)}{p(\vec{f} Me -)p(Me -)} \right)$ where $p(Me+)$ and $p(Me-)$ are the prior probabilities of methylated or un-methylated positions, respectively.” EX1007, 66:46-67:36. <i>See also supra</i> Sections XI.A.3-4, XI.A.7 (discussing elements 1.2, 1.3, and 1.6); EX1006, Abstract, 465-67, Fig. 3.
--	---

143. Accordingly, the Flusberg Publications in view of PacBio-BAM teach element 1.7.

144. Accordingly, the prior art renders claim 1 obvious.

B. Claim 2

145. Claim 2 depends from claim 1, reciting as follows:

The method of claim 1, wherein:

the input data structure is one input data structure of a plurality of input data structures,

the sample nucleic acid molecule is one sample nucleic acid molecule of a plurality of sample nucleic acid molecules,

the plurality of sample nucleic acid molecules are obtained from a biological sample of a subject, and

each input data structure corresponds to a respective window of nucleotides sequenced in a respective sample nucleic acid molecule of the plurality of sample nucleic acid molecules, and

the method further comprising:

receiving the plurality of input data structures,

inputting the plurality of input data structures into the model, and

determining, using the model, whether a modification is present in a nucleotide at a target location in the respective window of each input data structure.

This is a lengthy description of a simple concept: claim 2 is directed to performing each of elements 1.2, 1.3, 1.4, and 1.7 on input data structures from a plurality of sample nucleic acid molecules obtained from a biological sample of a subject.

(Claim 2 does not require the receiving step of element 1.1 or the training of elements 1.5 and 1.6 to be performed any additional times.) Claim 2 specifies that each input data structure corresponds to a window of its respective sample nucleic acid molecule, and each of the data structures is received and input into the model, which determines whether a modification is present at a respective target location in the window of that input data structure. The Flusberg Publications in view of PacBio-BAM disclose or teach each element of claim 2.

146. The Flusberg Publications teach determining whether a modification is present on a target nucleotide for each of a plurality of nucleic acid molecules from a subject's biological sample. *See, e.g.*, EX1007, 18:5-16 (discussing detecting "fluorescence signals from an array of thousands of ZMWs simultaneously"), 20:63-21:2 (discussing "highly parallel" analysis of entire methylome), 2:11-22 (discussing mapping of human methylome), 74:57-75:30 (application of metagenomics, including identification of analyzed organisms); EX1006, Fig. 3 (showing methylation determination for multiple different nucleic acid molecules), 463 (discussing methylation analysis of multiple molecules). A person of ordinary skill in the art would have understood that the purpose of developing a model like the ones taught by the Flusberg Publications is to detect modifications from multiple nucleic acid molecules, and in particular from multiple nucleic acid molecules derived from a biological sample of a subject (e.g., a subject whose methylome was to be sequenced). Indeed, a biological sample from a subject would typically contain a large number of DNA templates, and given Flusberg-338's disclosure of analyzing multiple ZMWs in parallel, each containing a different nucleic acid molecule, a person of ordinary skill in the art would find it natural to analyze the resulting plurality of pulse traces from the subject's different nucleic acid molecules. Doing this type of analysis of multiple nucleic acid molecules from a subject would provide the recited plurality of data

structures, each having a window around a respective target nucleotide, and each producing a respective determination of modification at the target location.

'794 patent	Prior Art
2. The method of claim 1, wherein: the input data structure is one input data structure of a plurality of input data structures, the sample nucleic acid molecule is one sample nucleic acid molecule of a plurality of sample nucleic acid molecules, the plurality of sample nucleic acid molecules are obtained from a biological sample of a subject, and each input data structure corresponds to a respective window of nucleotides sequenced in a respective sample nucleic acid molecule of the plurality of sample nucleic acid molecules, and the method further comprising: receiving the plurality of input data structures,	“The preferred direct methylation sequencing described herein may generally be carried out using single molecule real time sequencing systems, i.e., that illuminate and observe individual reaction complexes continuously over time, such as those developed for SMRT™ DNA sequencing (see, e.g., P. M. Lundquist, et al., <i>Optics Letters</i> 2008, 33, 1026, which is incorporated herein by reference in its entirety for all purposes). The foregoing SMRT™ sequencing instrument generally detects fluorescence signals from an array of thousands of ZMWs simultaneously, resulting in highly parallel operation. Each ZMW, separated from others by distances of a few micrometers, represents an isolated sequencing chamber.” EX1007, 18:5-16. “Carried out in a highly parallel manner, methylomes can be sequenced directly, with single base-pair resolution and high throughput.” EX1007, 20:63-21:2. <i>See also</i> EX1006, 463, Fig. 3; EX1007, 2:11-22, 74:57-75:30; <i>supra</i> Sections XI.A.3-5, XI.A.8.

inputting the plurality of input data structures into the model, and determining, using the model, whether a modification is present in a nucleotide at a target location in the respective window of each input data structure.	
---	--

147. Accordingly, it is my opinion that claim 2 would have been obvious over the Flusberg Publications in view of PacBio-BAM.

C. Claim 7

148. Claim 7 depends from claim 2 and further recites “wherein the modification is a methylation, the method further comprising: determining the modification is present at one or more nucleotides, and determining a clinically-relevant DNA fraction, a fetal methylation profile, a maternal methylation profile, a presence of an imprinting gene region, or a tissue of origin using the presence of the modification at one or more nucleotides.” Thus, claim 7 imposes two requirements: (1) the method must determine that the modification *is* present at one or more nucleotides (i.e., not all nucleotides turn out to be unmodified); and (2) one of several alternative actions must be taken using the presence of the modification at the one or more nucleotides. The Flusberg Publications in view of PacBio-BAM disclose or teach each element of claim 7.

149. For the first element of claim 7, the Flusberg Publications teach determining that the modification is present at one or more nucleotides. Indeed, this is the primary purpose of the Flusberg Publications. *See, e.g.*, EX1006, Abstract (describing “direct detection of modified nucleotides in the DNA template, including n6-methyladenine, 5-methylcytosine and 5-hydroxymethylcytosine”); EX1007, 66:46-67:4 (describing “a classifier for distinguishing methylated from unmethylated cytosines”), 3:47-53 (“detection of methylated bases within nucleic acid sequences using a real time direct detection of such methylated sites”). The normal use of the process I discuss above for claim 1, when applied to a typical biological sample of the type contemplated by the Flusberg Publications that would contain at least some detectable modifications, would result in determination that a modification is present at one or more nucleotides. *See supra* Section XI.A.8 (discussing modification detection for claim 1).

150. With regard to remaining element, the the Flusberg Publications teach at least one of the alternative actions recited in claim 7. For example, the Flusberg Publications teach identification of a tissue of origin based on the presence of methylation being detected in one or more nucleotides. Flusberg-338 teaches that “methylation status differs between tissue types.” EX1007, 2:11-16. Thus, a person of ordinary skill in the art would have understood that detected methylation could

be used to distinguish between tissues of origin for a sample—that is, by detecting methylation consistent with a first type of tissue but inconsistent with a second (based on the known differences between tissue types taught by Flusberg-338), a person of ordinary skill in the art would have been able to determine that the type of tissue from which the sample originated was the first, not the second. This reasoning is reflected in Flusberg-338, which teaches that methylation has been “associated with cancer” and hydroxymethylation is an epigenetic marker “identified in certain cell types in the brain” (EX1007, 1:55-2:10), so detection of methylation patterns could be used to identify a tissue of origin as cancerous or as containing such brain cells. Flusberg-338 also teaches that methylation may distinguish which organism was the origin of tissue, disclosing that one may distinguish between organisms having “very similar” genomes “by association with the known methylation patterns” of the organisms. EX1007, 7:1-7. Applying these examples to the technique taught by the Flusberg Publications in view of PacBio-BAM would result in “determining ... a tissue of origin using the presence of the modification at one or more nucleotides” as recited in claim 7.

'794 patent	Prior Art
7. The method of claim 2, wherein the modification is a methylation,	“We describe the direct detection of DNA methylation, without bisulfite conversion ... direct detection of modified nucleotides in the DNA template, including n6-methyladenine, 5-

the method further comprising: determining the modification is present at one or more nucleotides, and determining a clinically-relevant DNA fraction, a fetal methylation profile, a maternal methylation profile, a presence of an imprinting gene region, or a tissue of origin using the presence of the modification at one or more nucleotides.	methylcytosine and 5-hydroxymethylcytosine.” EX1006, Abstract. “The present invention is generally directed to the detection of modified nucleic acid sequences, and particularly the detection of methylated bases within nucleic acid sequences using a real time direct detection of such methylated sites.” EX1006, 3:47-53. “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.” EX1007, 66:46-67:4. “Recently, a new 5-hydroxymethylcytosine epigenetic marker has been identified in certain cell types in the brain, suggesting that it plays a role in epigenetic control of neuronal function.... In contrast to determining a human genome, mapping of the human methylome is a more complex task because the methylation status differs between tissue types....” EX1007, 1:55-2:16. “[I]f a particular organism’s genome is known to have a specific DNA methylation pattern that is distinct from another organism, despite their primary sequences being very similar, the kinetic signatures can be detected by SMRT™ sequencing and used to help identify the organism by association with the known methylation patterns....” EX1007, 7:1-7. <i>See also</i> EX1007, 7:1-30.
--	--

151. Accordingly, it is my opinion that claim 7 would have been obvious over the Flusberg Publications in view of PacBio-BAM.

D. Claim 10

152. Claim 10 recites the “method of claim 1, further comprising sequencing the sample nucleic acid molecule.” The Flusberg Publications both disclose sequencing the sample nucleic acid molecule, as such sequencing is the basis for detection of modifications. For example, Flusberg 388 teaches that “[m]ethylation sequencing on a SMRT™ Sequencing platform ... 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.” EX1007, 75:32-46; *see also id.*, 11:1-39, 60:45-67; EX1006, 461-63, 465, 466.

'794 patent	Prior Art
10. The method of claim 1, further comprising sequencing the sample nucleic acid molecule.	“Methylation sequencing on a SMRT™ Sequencing platform (see, e.g., P. M. Lundquist, et al., <i>supra</i>) 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.” EX1007, 75:32-46. “We sequenced these templates and then compared the average IPD in each of the methylated templates to the average IPD in the control template by computing their ratio at every template position.” EX1006, 462. <i>See also</i> EX1006, 461-63, 465, 466; EX1007, 11:1-39, 60:45-67.

153. Accordingly, it is my opinion that claim 10 would have been obvious over the Flusberg Publications in view of PacBio-BAM.

E. Claim 11

154. Claim 11 depends from claim 1 and further recites that “the model includes a machine learning model, a principal component analysis, a convolutional neural network, or a logistic regression.” The Flusberg Publications in view of PacBio-BAM disclose or teach this element.

155. With regard to the first alternative listed in claim 11, the ’794 patent specifies that “‘machine learning models’ may include models based on using sample data (e.g., training data) to make predictions on test data, and thus may include supervised learning,” and “may include statistical models.” EX1001, 16:34-40. In other words, a “machine learning model” includes a model that uses training data to predict test data, including a statistical model for making such predictions. Flusberg-338 discloses this type of machine learning model. *See* EX1007, 66:46-67:36 (describing machine learning, including “developing a classifier on a training set”), 67:64-68:5 (“standard statistical classification algorithms”). The Flusberg Publications also disclose other options recited in claim 11, such as principal component analysis (EX1006, Fig. 3, 463, 466; EX1007, 67:27-30, 69:36-47) and regression (EX1007, 67:64-68:5), which a person of ordinary skill in the art would recognize would include an option of various well-

known types of regression, including logistic regression. *See* EX1017, 23 (“logistic regression” is one of the “standard machine learning methods used in genomics”). Flusberg-338 also discloses that the model may be a neural network. EX1007, 67:64-68:5. A person of ordinary skill in the art would have recognized that two primary types of neural network are convolutional neural networks and recurrent neural networks, as I discussed briefly above for claim 1. *See supra* Section XI.A.7; EX1017, 1-2 (describing convolutional neural networks and recurrent neural networks as known neural network architectures appropriate for genomics applications); EX1035, Abstract (using convolutional neural network to predict methylation sites). A person of ordinary skill in the art would have thus viewed a convolutional neural network as a natural, predictable option to implement a model for detecting modifications as taught by the Flusberg Publications.

'794 patent	Prior Art
11. The method of claim 1, wherein the model includes a machine learning model, a principal component analysis, a convolutional neural network, or a logistic regression.	“To this end, we 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....” EX1006, 463. “Standard principal component analysis was carried out....” EX1006, 466 (citation omitted). “In certain embodiments, the major component in a principal components analysis can serve as a better weighted combination of the most important features.” EX1007, 67:27-30.

	“In addition, the general classifier approach suggested in steps 2+3 can be implemented with many standard statistical classification algorithms, i.e. linear discriminant analysis, multi-dimensional regression, kernel methods, classification and regression trees, neural networks, and support vector machines.” EX1007, 67:64-68:5. <i>See also</i> EX1006, Fig. 3; EX1007, 66:46-67:36, 69:36-47.
--	---

156. Accordingly, it is my opinion that claim 11 would have been obvious over the Flusberg Publications in view of PacBio-BAM.

F. Claim 12

157. Claim 12 depends from claim 1 and further recites that “the window of nucleotides corresponding to the input data structure comprises nucleotides on a first strand of the sample nucleic acid molecule and nucleotides on a second strand of the sample nucleic acid molecule, and the input data structure further comprises for each nucleotide within the window a value of a strand property, the strand property indicating the nucleotide being present on either the first strand or the second strand.” The Flusberg Publications in view of PacBio-BAM disclose or teach each element of claim 12.

158. For example, Flusberg-338 teaches that a complementary pair of strands may be analyzed to detect methylation at a CG sequence which has a complementary CG sequence that may be jointly methylated. *See* EX1007, 68:5-21 (also noting that GATC sequences, related to adenosine methylation, have identical

reverse complements as well). Flusberg-338 explains that a model can take both strands into account to improve accuracy of methylation detection, because methylation at a locus on one strand is predictive of methylation of the adjoining locus on the other strand. *See id.* The complementary nucleotides on each strand would each be nucleotides in a window about the nucleotide of interest (i.e., the C at either of the complementary CG loci), and a person of ordinary skill in the art would understand that indicating which nucleotides were on each strand would aid in performing this analysis of complementary strands. Further, this type of indication would already be present in a PacBio BAM file, as PacBio-BAM teaches that such a file contains indications identifying which nucleotides are on each complementary strand. *See* EX1011, 6 (indicating “forward” or “reverse” pass in “subread local context”).

'794 patent	Prior Art
12. The method of claim 1, wherein: the window of nucleotides corresponding to the input data structure comprises nucleotides on a first strand of the sample nucleic acid molecule and nucleotides on a second strand of the sample	“The approach can also incorporate data from multiple strands of a duplex template. For example, because the CG sequence for cytosine methylation and the GATC sequence for adenosine methylation is the same on the reverse complement strand, these bases can be methylated on both complementary strands. If the general statistical distribution for the fraction of sites that are hemi-methylated vs. fully methylated is known, then information regarding IPD or other metrics gained from the complementary strand can be used to increase the accuracy with which a call is made on a particular strand. For example, if after analyzing each strand separately it is

nucleic acid molecule, and the input data structure further comprises for each nucleotide within the window a value of a strand property, the strand property indicating the nucleotide being present on either the first strand or the second strand.	concluded that there is a 95% chance that stand A is methylated and a 55% chance that complementary strand B is methylated, but it is known that there is a 80% chance that if one strand is methylated then so is the other, then the confidence in calling strand B as methylated is increased.” EX1007, 68:5-21. <i>See also</i> EX1011, 1-8.
---	---

159. Accordingly, it is my opinion that claim 12 would have been obvious over the Flusberg Publications in view of PacBio-BAM.

G. Claims 14 and 17

160. Claims 14 and 17 each depend from claim 1. Claim 14 further recites that “the nucleotides within the window are determined using a circular consensus sequence and without alignment of the sequenced nucleotides to a reference genome.” Claim 17 further recites that “nucleotides within the window are determined without using a circular consensus sequence and without alignment of the sequenced nucleotides to a reference genome.” Thus, claims 14 and 17 both recite that the nucleotides are determined “without alignment of the sequenced nucleotides to a reference genome,” and they further specify that circular consensus sequencing either *is* (claim 14) or *is not* (claim 17) used. The Flusberg Publications in view of PacBio-BAM disclose or teach each element of claims 14 and 17.

161. Beginning with the common element of claims 14 and 17, PacBio-BAM teaches that a BAM file may optionally be either aligned or unaligned to a genome. EX1011, 1-2. Flusberg also discloses that synthetic sequences may be tested for modifications. EX1006, 462, Fig. 4; *see also* EX1007, 13:52-58, 75:33-46. A person of ordinary skill in the art would understand that such sequences cannot be aligned to a reference genome, as a synthetic sequence has no *genome* to align to: it would be determined from a selected artificial nucleotide sequence, rather than from the genome of an organism, so it would make sense in such examples to forego alignment to a genome. A person of ordinary skill in the art would thus understand that while alignment to a genome may be helpful in some contexts (e.g., when dealing with biological samples from subjects with a known genome), using an unaligned sequence may be helpful or even necessary in others (as when the nucleic acid strand is artificial in origin or a subject's genome is unknown or not easily mapped).

162. Regarding circular consensus sequencing, a person of ordinary skill in the art would have known that circular consensus sequencing would be an option, not a universal requirement, with some situations calling for circular consensus sequencing and others better served by not employing it. Flusberg-338 discloses that its modification detection can be performed on “double-stranded linear nucleic acid [or] double-stranded circular nucleic acid.” EX1007, 6:56-61; *see also, e.g.,*

id., 9:35-41 (“Preferably, the processing is a sequencing reaction (e.g., a single-molecule sequencing reaction), and where the template is a closed circular template and the polymerase is capable of strand displacement, the processing can comprise rolling-circle replication of the template, which can generate redundant sequence data for the template.”), 28:53-57 (redundant sequencing not required for methylation detection). Flusberg-338 thus discloses that circular consensus sequencing is optional, but that sequencing of a linear (i.e., non-circular) nucleic acid may also be performed, and that methylation detection may be performed without redundant (e.g., consensus) sequencing. A person of ordinary skill in the art would thus consider performing circular consensus sequencing (as in claim 17) and not performing it (as in claim 14) as known, acceptable options to be freely chosen according to the particular application.

'794 patent	Prior Art
14. The method of claim 1, wherein the nucleotides within the window are determined using a circular consensus sequence and without alignment of the sequenced nucleotides to a reference genome.	“To test this hypothesis, we designed several synthetic DNA templates that were identical except for their methylation status at specific sites.” EX1006, 462. “The template can comprise, e.g., single-stranded linear nucleic acid, single-stranded circular nucleic acid, double-stranded linear nucleic acid, double-stranded circular nucleic acid, or a combination thereof.” EX1007, 6:56-61.
17. The method of claim 1, wherein nucleotides within the window are	“Preferably, the processing is a sequencing reaction (e.g., a single-molecule sequencing reaction), and where the template is a closed circular template and the polymerase is capable of strand displacement, the

determined without using a circular consensus sequence and without alignment of the sequenced nucleotides to a reference genome.	processing can comprise rolling-circle replication of the template, which can generate redundant sequence data for the template.” EX1007, 9:35-41. <i>See also</i> EX1006, Fig. 4; EX1007, 13:52-58, 28:53-57, 75:33-46; EX1011, 1-8.
--	--

163. Accordingly, it is my opinion that claims 14 and 17 would have been obvious over the Flusberg Publications in view of PacBio-BAM.

H. Claim 15

164. Claim 15 depends from claim 1 and further recites that “each nucleotide within the window is enriched or filtered.” I understand that by reciting that each nucleotide be enriched or filtered, claim 15 encompasses not only processes that both enrich some nucleotides and filter some nucleotides, but also processes that enrich all nucleotides (even if none are filtered), as well as processes that filter all nucleotides (even if none are enriched). In this section, I address the alternative in which all nucleotides are enriched.

165. Flusberg-338 teaches this element. For example, Flusberg-338 teaches that “[e]nrichment strategies are beneficial both for reducing experimental variation and increasing efficiency of identifying modifications within a nucleic acid sample,” as such strategies result in nucleic acid samples that are “less complex and more enriched for fragments of interest than was the original nucleic acid sample.” EX1007, 60:22-32. Flusberg-338 provides detailed explanations of how nucleic acid molecules may be enriched to enhance modification detection, as

by selectively binding nucleic acid molecules to binding agents to allow modifications or fragments of interest to be preferentially sequenced. *See* EX1007, 59:45-61:5. By performing such enrichment to identify target fragments, the nucleotides composing the selected fragments all would be enriched (by being in a fragment selected via the enrichment process). Thus, Flusberg-338 expressly teaches that the nucleotides within the window may be enriched.

'794 patent	Prior Art
15. The method of claim 1, wherein each nucleotide within the window is enriched or filtered.	“Enrichment strategies are beneficial, both for reducing experimental variation and increasing efficiency of identifying modifications within a nucleic acid sample. For example, in certain preferred embodiments a nucleic acid sample is exposed to a binding agent that specifically binds a modification of interest, the nucleic acid sample is fragmented, and fragments comprising the modification are retained by virtue of their association with the binding agent. In doing so, the resulting nucleic acid sample is less complex and more enriched for fragments of interest than was the original nucleic acid sample.” EX1007, 60:22-32. <i>See also</i> EX1007, 59:45-61:5.

166. Accordingly, it is my opinion that claim 15 would have been obvious over the Flusberg Publications in view of PacBio-BAM.

I. Claim 18

167. Claim 18 depends from claim 1 and further recites that “the optical signal is a fluorescence signal from a dye-labeled nucleotide.” The Flusberg Publications teach this element. For example, Flusberg teaches that the nucleotides

generating the optical signals may each be associated with a dye, such as “Alexa Fluor 555-dT, Alexa Fluor 568-dG, Alexa Fluor 647-dA and Cy5.5-dC.” EX1006, 466; *see also id.*, Abstract, 461-62. Flusberg-338 echoes this teaching, noting that its nucleotides have “a detectable label (e.g., comprising a fluorescent dye) is attached to the terminal phosphate rather than the base.” EX1007, 21:42-45; *see also id.*, 22:1-24, 22:57-67.

'794 patent	Prior Art
18. The method of claim 1, wherein the optical signal is a fluorescence signal from a dye-labeled nucleotide.	“The second important component is phospholinked nucleotides for which a detectable label (e.g., comprising a fluorescent dye) is attached to the terminal phosphate rather than the base (Fig. 1)...” EX1007, 21:42-45. “Additional modifications included permuting the nucleobases associated with each dye to the following configuration: Alexa Fluor 555-dT, Alexa Fluor 568-dG, Alexa Fluor 647-dA and Cy5.5-dC.” EX1006, 466. <i>See also</i> EX1006, Abstract, 461-62; EX1007, 22:1-24, 22:57-67.

168. Accordingly, it is my opinion that claim 18 would have been obvious over the Flusberg Publications in view of PacBio-BAM.

J. Claim 19

169. Claim 19 depends from claim 1 and further recites that “each window associated with the first plurality of data structures comprises 4 consecutive nucleotides on a first strand of each first nucleic acid molecule.” This applies to the

data structures used in training, not necessarily to the data structure used for the determination recited in claim 1, requiring that the window size for those data structures be at least 4 consecutive nucleotides. The Flusberg Publications disclose or teach this element.

170. For example, Flusberg-338 teaches a window may vary in size, giving an example of a 7-nucleotide window of consecutive nucleotides, from the nucleotide labeled $i-3$ to the nucleotide labeled $i+3$. EX1007, 66:46-67:4 (“[A] local context extending 3 pulses to the left and right of the pulse of interest is assumed, but this context size is flexible....”). Flusberg likewise teaches using a window comprising 4 consecutive nucleotides. *See* EX1006, Supplementary Table 1 (principal component weights computed for a 9-nucleotide window). A person of ordinary skill in the art would have viewed it as natural to use window sizes greater than 4 consecutive nucleotides, as disclosed by the Flusberg Publications, in the data structures used to train a model for detecting modifications.

'794 patent	Prior Art
19. The method of claim 1, wherein each window associated with the first plurality of data structures comprises 4 consecutive nucleotides on a first strand of each first nucleic acid molecule.	“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. <i>See also</i> EX1006, Supplementary Table 1.

171. Accordingly, it is my opinion that claim 19 would have been obvious over the Flusberg Publications in view of PacBio-BAM.

172. Accordingly, it is my opinion that claims 1-2, 7, 10-12, 14-15, and 17-19 would have been obvious in view of the Flusberg Publications and PacBio-BAM.

XII. GROUND 2 – CLAIMS 3-6 AND 9 ARE OBVIOUS AS ILLUSTRATED BY THE FLUSBERG PUBLICATIONS, PACBIO-BAM, AND CLEMENT

173. As I explain in greater detail below, the Flusberg Publications, PacBio-BAM, and Clement render claims 3-6 and 9 obvious. Claims 3-6 depend from claim 2 and relate to classifications of a disorder such as cancer based on the presence of modifications determined according to the methods described in claims 1 and 2. Claim 9 also depends from claim 2 and recites elements related to alignment to a genome and use of a cutoff number, both of which are steps useful when performing the type of analysis typically employed for detecting the kind of disorders addressed in claims 3-6. The '794 patent defines “classification” as referring “to any number(s) or other characters(s) that are associated with a particular property of a sample.” EX1001, 14:65-15:5. This includes binary classifications as well as more subtle classifications (such as classifications of different levels of disease). *See id.* For example classifications for cancer may indicate a “level of cancer,” which may include any of numerous characteristics,

such as stage, size, metastasis, response to treatment, recurrence, or other measures. EX1001, 15:25-45.

174. The Flusberg Publications teach that “DNA methylation ... has been associated with cancer and other disease processes,” but they do not describe in detail how such disorders can be classified based on methylation patterns. EX1007, 1:56-61; *see* EX1006, 461. These details are provided by Clement. Clement teaches “methods of determining a cancer treatment prognosis” based on “evaluating epigenetic and genetic changes within a tumor sample.” EX1018, Abstract. The epigenetic evaluation contemplated by Clement includes classifying patterns of methylation in nucleic acid molecules obtained from subjects. EX1018, ¶¶9, 78, 198. In light of the Flusberg Publications’ teaching that methylation is associated with cancer and Clement’s teaching of how that association can be used to classify cancer, a person of ordinary skill in the art would have been motivated to classify cancer in a patient based on a pattern of methylation detected as taught by the Flusberg Publications and PacBio-BAM to aid in assessing cancer prognosis.

175. A person of ordinary skill in the art would have a reasonable expectation of success in combining Clement with the Flusberg Publications and PacBio-BAM. The combination would be predictable and routine, as it would merely use SMRT-detected methylation as taught by Flusberg and PacBio-BAM to

classify cancer as taught by Clement. Each element would perform its ordinary function in the combination, with methylation patterns being detected as taught by the Flusberg Publications and PacBio-BAM, and those patterns being used as input data to classify cancer as taught by Clement. The combination of these references would thus be a combination of elements according to known methods to yield predictable results.

A. Claims 3 and 4

176. Claim 3 depends from claim 2 and recites that the method further comprises “determining the modification is present at one or more nucleotides, and determining a classification of a disorder using the presence of the modification at one or more nucleotides.” Claim 4 depends from claim 3 and further recites that “the disorder comprises cancer.” The Flusberg Publications in view of PacBio-BAM and Clement disclose or teach each element of claims 3 and 4.

177. Regarding the first element of claim 3, the Flusberg Publications teach “determining the modification is present at one or more nucleotides,” as recited in claim 3. Indeed, this portion of claim 3 recites the same limitation as I addressed above for claim 7, and it is met for claim 3 for similar reasons. *See supra* Section XI.C (discussing this element for claim 7); EX1006, Abstract; EX1007, 3:47-53, 66:46-67:4.

178. Regarding the second element of claim 3, the Flusberg Publications and Clement teach determining a classification of a disorder comprising cancer, using the recited presence of the modification. As I discussed above, the Flusberg Publications teach that changes in methylation patterns are associated with cancer and other diseases. EX1006, 461; EX1007, 1:56-61. Clement further teaches that analyzing methylation patterns allows classification of a disease such as cancer. For example, Clement teaches that a methylation pattern for a subject may be classified as consistent (“concordant”) or inconsistent (“discordant”). EX1018, ¶78 (“The terms ‘locally disordered methylation’ and ‘discordant methylation’ are used interchangeably and refer [to] alterations of CpG methylation patterns over a short genetic distance or within a genomic feature. Typically, short-range concordance is expected to be very high in non-disease states, as DNA methylation generally changes by feature (e.g., a specific gene promoter, or a CG island) rather than by individual CpG. These terms may also refer, in a stricter sense, to the concordance status of CpGs on the same sequencing read.”). Discordant or inconsistent methylation is associated with malignancy in a cancer patient, with inconsistent methylation over a sequence or across multiple genes indicating a “subject is more likely to (1) develop resistance to an antitumor agent; (2) relapse after treatment with an antitumor agent; (3) develop a metastatic tumor; or (4) any combination thereof.” EX1018, ¶¶9, 198. Applying these principles as taught by Clement in

combination with the methylation determination for nucleic acids from a patient's biological sample, as taught by Flusberg Publications and PacBio-BAM, the patient would be classified as likely having malignant cancer where a discordant methylation pattern was detected. This would be "determining a classification of a disorder using the presence of the modification at one or more nucleotides," with the particular disorder comprising cancer (e.g., presence or absence of malignant cancer, cancer likely to metastasize, relapse, or develop resistance, etc.).

'794 patent	Prior Art
3. The method of claim 2, further comprising: determining the modification is present at one or more nucleotides, and determining a classification of a disorder using the presence of the modification at one or more nucleotides. 4. The method of claim 3, wherein the disorder comprises cancer.	"In addition, DNA methylation is known to be essential for mammalian development and has been associated with cancer and other disease processes." EX1007, 1:56-61. "Changes in mC patterns have a crucial role in development and have been associated with cancer and other diseases." EX1006, 461. "The presence or prevalence of inconsistent methylation status along the sequences or across the multiple gene copies may indicate that the subject is more likely to (1) develop resistance to an antitumor agent; (2) relapse after treatment with an antitumor agent; (3) develop a metastatic tumor; or (4) any combination thereof." EX1018, ¶9. "These results support the idea that locally disordered methylation is a general property of the malignant process." EX1018, ¶198. "The terms 'locally disordered methylation' and 'discordant methylation' are used interchangeably and refer alterations of CpG methylation patterns over

	a short genetic distance or within a genomic feature.” EX1018, ¶78. <i>See also supra</i> Section XI.C (discussing claim 7); EX1006, Abstract; EX1007, 3:47-53, 66:46-67:4.
--	---

179. Accordingly, it is my opinion that claims 3 and 4 would have been obvious over the Flusberg Publications in view of PacBio-BAM and Clement.

B. Claim 5

180. Claim 5 depends from claim 4 and further recites “determining that the classification of the disorder is that the subject has the disorder, and treating the subject for the disorder by chemotherapy, radiation, or surgery.” The Flusberg Publications in view of PacBio-BAM and Clement disclose or teach each element of claim 5.

181. Chemotherapy, radiation, and surgery are all long-established treatments for cancer. As I discussed above for claims 3 and 4, Clement teaches determining that a patient has “the disorder” (i.e., a malignant cancer more likely to develop resistance, relapse, or metastasize) based on a detected discordant methylation pattern. EX1018, ¶¶9, 198. Clement further teaches that in such a case, a patient may be treated with “surgery or radiation,” or by “altering treatment with an antitumor agent.” EX1018, ¶19. As Clement explains, it uses the terms “antitumor agent” and “chemotherapeutic agent” interchangeably, so the alteration of treatment “with an antitumor agent” refers to a chemotherapy treatment.

EX1018, ¶70. Clement thus teaches that if the patient is determined to have the disorder, they should be treated with chemotherapy, radiation, or surgery.

'794 patent	Prior Art
5. The method of claim 4, further comprising: determining that the classification of the disorder is that the subject has the disorder, and treating the subject for the disorder by chemotherapy, radiation, or surgery.	“Optionally, the method comprises ceasing or altering treatment with an antitumor agent, or initiating a non-chemotherapeutic treatment (e.g., surgery or radiation) if a high level of inconsistencies in methylation status is identified.” EX1018, ¶19. “The terms ‘antitumor agent’ and ‘chemotherapeutic agent’ are used interchangeably herein and refer to an agent for the treatment of cancer.” EX1018, ¶70. <i>See also</i> EX1018, ¶¶9, 198.

182. Accordingly, it is my opinion that claim 5 would have been obvious over the Flusberg Publications in view of PacBio-BAM and Clement.

C. Claim 6

183. Claim 6 depends from claim 3 and further recites that “determining the classification of the disorder uses the number of modifications or the sites of the modifications.” Clement teaches this element.

184. As I discussed above for claims 3 and 4, Clement teaches that discordant methylation patterns indicate likely malignant cancer (e.g., cancer more likely to develop resistance, relapse, or metastasize), whereas concordant patterns indicate less likely malignancy. EX1018, ¶¶9, 198. A discordant pattern is one where methylation is “locally disordered,” in contrast to a “concordant” pattern in

which “all CpGs contained within one sequencing read are uniformly methylated or uniformly unmethylated.” EX1018, ¶78. The concordant/discordant determination relates to the number or sites of modifications because it compares the number of methylations in a local sequence region to the number of CpG sites, with concordance corresponding to all or no CpG sites in the region methylated and discordant corresponding to some but not all local CpG sites methylated.

'794 patent	Prior Art
6. The method of claim 3, wherein determining the classification of the disorder uses the number of modifications or the sites of the modifications.	“The terms “locally disordered methylation” and “discordant methylation” are used interchangeably and refer alterations of CpG methylation patterns over a short genetic distance or within a genomic feature. ... If all CpGs contained within one sequencing read are uniformly methylated or uniformly unmethylated, the read is classified as concordantly methylated. Otherwise the read is classified as discordantly methylated.” EX1018, ¶78. <i>See also</i> EX1018, ¶¶9, 198, Fig. 3.

185. Accordingly, it is my opinion that claim 6 would have been obvious over the Flusberg Publications in view of PacBio-BAM and Clement.

A. Claim 9

186. Claim 9 depends from claim 2 and further recites the following:

the plurality of sample nucleic acid molecules align to a plurality of genomic regions,

for each genomic region of the plurality of genomic regions:

a number of sample nucleic acid molecules is aligned to the genomic region,

the number of sample nucleic acid molecules is greater than a cutoff number.

The Flusberg Publications in view of PacBio-BAM and Clement disclose or teach each element of claim 9.

187. PacBio-BAM teaches that the sequences in a BAM file may be aligned to a reference genome. EX1011, 1-2. The Flusberg Publications agree. *See* EX1006, 461 (“align to the reference genome”), 466 (alignment using “Smith-Waterman algorithm”); EX1007, 50:30-33 (discussing “align[ing] and map[ping] to a known sequence, e.g., from a genomic sequence database”), 23:32-36 (“comparison to a reference sequence”). The Flusberg Publications further contemplate aligning multiple nucleic acid molecules’ sequences to multiple genomic regions, including performing genome-wide mapping. EX1007, 2:11-22 (discussing “determination of genome-wide methylation patterns”), 13:24-40 (discussing mapping methylation “across an entire genome”); EX1006, 461 (proposing “mapping genome-wide mC patterns”).

188. Clement similarly teaches aligning to a reference genome, and it further teaches analyzing a “Proportion of Discordant Reads (PDR)” to assess methylation discordance “across the entire genome or for a specified genomic region.” EX1018, ¶235. This measure is given by “averaging the values of individual CpGs, as calculated for all CpGs within the region of interest with read depth greater than 10.” *Id.* By requiring read depth of greater than 10, Clement is

setting a cutoff where at least 10 nucleic acids must overlap at a CpG site in order to include the region containing that CpG as part of that measure. This use of a cutoff based on read depth was a well-known method in bioinformatics that was useful to increase confidence in the reliability of a bioinformatic measure. By setting a read depth cutoff of 10 or higher, Clement reduces the likelihood that a given methylation pattern is a result of random noise. *See also* EX1018, ¶¶55 (“at least 4 CpG per read and read depth >10”), 56 (“read depth of 30”), 58 (“sufficient read depth (>40)”), 249 (“30× coverage depth”).

189. One of the figures provided by Clement provides an intuitive picture that helps to visualize why setting a read depth cutoff would aid in distinguishing true methylation discordance from a false signal generated by random noise. Fig. 3E, below, shows how discordant methylation patterns may be seen based on the number of reads, n , at each of a plurality of CpG sites in a region:

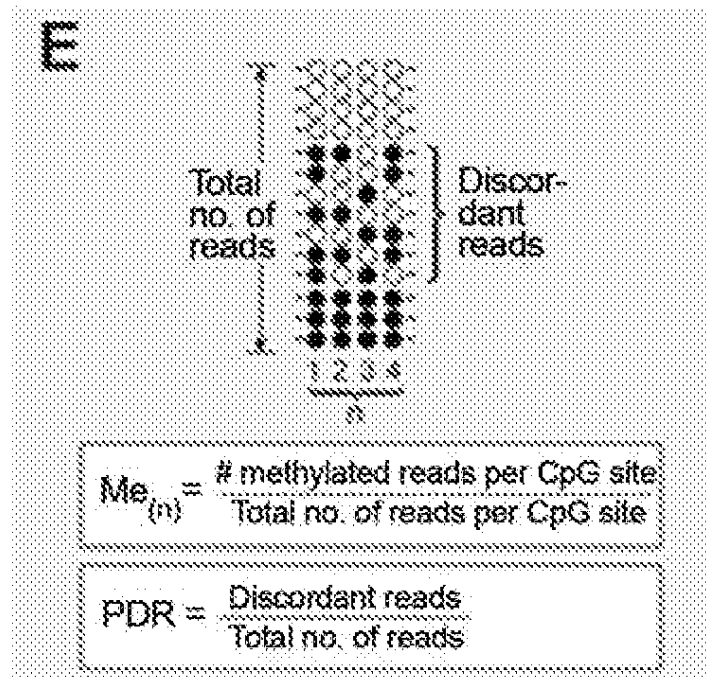


FIG. 3E

EX1018, Fig. 3E. As shown in the figure, the existence of discordant reads may be observed by identifying regions where CpG locations are varyingly methylated and unmethylated. Logically, no discordance can be detected at a site if only one nucleic acid covering that location is queried for methylation. Likewise, if only a small number of nucleic acid molecules were analyzed, a discordant region might be found to be concordant by chance. Likewise, errors in methylation detection might make a concordant region appear discordant unless a large enough number of nucleic acid molecules were measured to outweigh the measurement error.

190. In light of Clement’s teachings, a person of ordinary skill in the art have been motivated to apply a cutoff number when applying the methylation detection techniques taught by the Flusberg Publications and PacBio-BAM. In this combination, for each given genomic region being analyzed (e.g., for each CpG location), the number of distinct nucleic acid molecules analyzed for that region would be at least the cutoff number. This would be a straightforward application of a known technique to produce a predictable result: an increase in confidence that patterns detected represent real signal and not mere noise.

'794 patent	Prior Art
9. The method of claim 2, wherein: the plurality of sample nucleic acid molecules align to a plurality of genomic regions, for each genomic region of the plurality of genomic regions: a number of sample nucleic acid molecules is aligned to the genomic region, the number of sample nucleic acid molecules is greater than a cutoff number.	“The BAM format is a binary, compressed, record-oriented container format for raw or aligned sequence reads.” EX1011, 1. “Pulse calling, which used a threshold algorithm on the dye-weighted intensities of fluorescence emissions, and read alignments, achieved using a Smith-Waterman algorithm, have been described previously.” EX1006, 466 (citation omitted). “In fact, it is preferable to sequence a long template because a long sequencing read is generally easier to align and map to a known sequence, e.g., from a genomic sequence database.” EX1007, 50:30-33. “After reads are aligned to the reference genome, the methylation state of each CpG on a read is determined. If all the CpGs on a specific read are methylated, or all of the CpGs on a read are unmethylated, the read is classified as concordant; otherwise it is classified as discordant. At each CpG, the PDR is equal to the number of discordant reads that cover that location divided by the total number of

	reads that cover that location (FIG. 3E). The PDR across the entire genome or for a specified genomic region is given by averaging the values of individual CpGs, as calculated for all CpGs within the region of interest with read depth greater than 10 reads and that are covered by reads that contain at least 4 CpGs.” EX1018, ¶235. <i>See also</i> EX1006, 461; EX1007, 2:11-22, 13:24-40, 23:32-36, 31:56-59; EX1011, 1-2; EX1018, ¶¶55-56, 58, 249, Fig. 3E.
--	---

191. Accordingly, it is my opinion that claim 9 would have been obvious over the Flusberg Publications in view of PacBio-BAM and Clement.

XIII. GROUND 3 – CLAIMS 8 AND 15-16 ARE OBVIOUS AS ILLUSTRATED BY THE FLUSBERG PUBLICATIONS, PACBIO-BAM, AND THE TEMPLATE GUIDE

192. As I discuss in detail below, claims 8 and 15-16 would have been obvious over the Flusberg Publications in view of PacBio-BAM and the PacBio Template Guide. Claims 8 and 15-16 include limitations related to filtering DNA molecules, such as by size selection. As I discussed for claim 1, the Flusberg Publications teach using Pacific Biosciences’ SMRT sequencing to detect methylation. EX1006, Abstract; EX1007, 18:5-16. The PacBio Template Guide teaches the standard procedure for preparing DNA templates for PacBio sequencing, such as SMRT sequencing. EX1019, 2. These procedures include size selection steps. *E.g.*, EX1019, 24, 35. A person of ordinary skill in the art would have had reason to filter nucleic acid molecules using the size selection processes in the PacBio Template Guide because the Template Guide teaches the normal

procedures to prepare molecules for PacBio sequencing and “highly recommends” using size selection. EX1019, 35.

193. A person of ordinary skill in the art would have a reasonable expectation of success in combining these references. A person of ordinary skill in the art would have understood that PacBio SMRT sequencing would typically be performed in conjunction with the template preparation steps, such as size-selection, that the PacBio Template Guide teaches. Thus, the combination would merely involve combining elements according to known methods to yield routine, predictable results.

A. Claim 8

194. Claim 8 depends from claim 2 and further recites that “each sample nucleic acid molecule of the plurality of sample nucleic acid molecules has a size greater than a cutoff size.” The ’794 patent discloses that a “cutoff” refers to a “predetermined number,” and, for example, “a cutoff size can refer to a size above which fragments are excluded.” EX1001, 15:6-8. Claim 8 recites a cutoff in the opposite direction, with a plurality of nucleic acid molecules being above a given size. Claim 8 thus resembles the statement in the ’794 patent’s specification that each “sample nucleic acid molecule of the plurality of sample nucleic acid molecules may have a size greater than a cutoff size.” EX1001, 91:32-42 (listing numerous possible cutoff sizes, ranging from 100 bp to 1 million bp). I understand

that Patent Owner has argued that applying a process that includes molecules above a given size while excluding molecules below that size meets claim 8’s “cutoff size” limitation. The PacBio Template Guide teaches this element.

195. The PacBio Template Guide teaches a process to prepare nucleic acid templates such as double-stranded DNA molecules to be used in PacBio single molecule sequencing. *See* EX1019, 3. In one aspect of this process, the PacBio Template Guide teaches that it is “highly recommended to perform a size-selection step” when sequencing large libraries to avoid loading bias in ZMW wells. EX1019, 35. In one example, size selection is performed with AMPure beads to remove “ligation products smaller than 0.4 kb.” EX1019, 24. AMPure beads selectively bind DNA fragments based on their size, allowing DNA fragments to be separated based on whether they are above or below the cutoff size. The size selection step disclosed by the PacBio Template Guide excludes nucleic acid molecules below a cutoff size (e.g., 400 bases), leaving a remaining plurality of nucleic acid molecules each having a size above the cutoff. This step thus meets the limitations of claim 8.

'794 patent	Prior Art
8. The method of claim 2, wherein each sample nucleic acid molecule of the plurality of sample nucleic acid molecules	“When sequencing large insert libraries (e.g. 20 kb library) for generating long read lengths, it is highly recommended to perform a size-selection step to eliminate short SMRTbell templates that will preferentially load.” EX1019, 35.

has a size greater than a cutoff size.	“Perform all purification steps at room temperature to adequately remove enzymes (exonucleases, ligases, etc.) and ligation products smaller than 0.4 kb (e.g., adapter dimers).” EX1019, 24. <i>See also</i> EX1019, 2-3, Fig. 11.
--	---

196. Accordingly, it is my opinion that claim 8 would have been obvious over the Flusberg Publications in view of PacBio-BAM and the PacBio Template Guide.

B. Claims 15 and 16

197. Claim 15 depends from claim 1 and further recites “wherein each nucleotide within the window is enriched or filtered.” Claim 16 depends from claim 15 and further characterizes how the nucleotides are either enriched or filtered. Regarding filtering, claim 16 specifies that the nucleotides are “filtered by: selecting double-stranded DNA molecules having a size with a size range.” Read together, these claims indicate that the “nucleotides within the window” may be each “filtered” by selecting double-stranded DNA molecules having a size within a size range. The PacBio Template Guide teaches this element of both claims 15 and 16.

198. As I discussed above for claim 8, the PacBio Template Guide teaches size selection for double-stranded DNA molecules. *See* EX1019, 3, 24, 35. Selecting double-stranded DNA molecules with sizes above a cutoff constitutes selecting double-stranded DNA molecules having a size with a size range. In

particular, the size range is greater than the cutoff (e.g., if a cutoff of 400 bp were used, then the size range would be a range bounded on the bottom by 400 bp and with no upper limit). In another way of meeting this size range, Template Guide further teaches that it is “highly recommended to select amplicons of the same size (+/-10%) to minimize loading bias.” EX1019, 35. A size range of +/-10% would be a range with both upper and lower limits (ranging from 0.9 to 1.1 times the value chosen for the center of the range). Thus, selecting amplicons (i.e., double stranded DNA molecules) within a +/-10% size range would also meet the filtering and size range limitations of claims 15 and 16.

'794 patent	Prior Art
15. The method of claim 1, wherein each nucleotide within the window is enriched or filtered. 16. The method of claim 15, wherein each nucleotide within the window is enriched by: cutting a double-stranded DNA molecule using a Cas9 complex to form a cut double-stranded DNA molecule, and ligating a hairpin adaptor onto an end of the cut double-stranded DNA molecule, or	“Loading of SMRTbell templates into ZMWs is size dependent. Small inserts load better than large inserts. This is particularly important when sequencing different PCR amplicon sizes. It is highly recommended to select amplicons of the same size (+/- 10%) to minimize loading bias.” EX1019, 35. <i>See also</i> EX1019, 2-3, 24, 35, Fig. 11; <i>supra</i> Section XIII.A (discussing claim 8).

filtered by: selecting double-stranded DNA molecules having a size with a size range.	
--	--

199. Accordingly, it is my opinion that claims 15 and 16 would have been obvious over the Flusberg Publications in view of PacBio-BAM and the PacBio Template Guide.

XIV. GROUND 4 – CLAIMS 13 AND 15-16 ARE OBVIOUS AS ILLUSTRATED BY THE FLUSBERG PUBLICATIONS, PACBIO-BAM, AND TSAI

200. As I discuss in detail below, claims 13 and 15-16 would have been obvious over the Flusberg Publications, PacBio-BAM, and Tsai. Claims 13 and 15-16 recite steps related to enriching nucleotides by using a Cas9 complex and hairpin adapters to form circular DNA molecules. The Flusberg Publications teach ligating hairpin adapters to form circular DNA molecules, but they do not expressly teach cutting DNA using a Cas9 complex. EX1007, 60:54-59 (teaching use of hairpin adapters for enrichment), EX1006, 463 (teaching ligation of hairpin adapters). Tsai is a PacBio patent publication teaching cutting double-stranded DNA templates with a Cas9 complex and ligating hairpin adapters to form circular DNA molecules for SMRT sequencing. EX1020, ¶¶7, 41. Tsai further suggests using its techniques to prepare DNA for detection of base modifications “by monitoring the kinetics of the sequencing reaction,” citing to the PCT publication of Flusberg. EX1020, ¶23 (citing WO2012/065043, EX1024). Tsai teaches that

using its Cas9 complexes to prepare DNA templates for sequencing allows targeting of “regions having sites that may have epigenetic modifications” that are related to various disorders. EX1020, ¶102. In light of the combined teachings of Tsai, the Flusberg Publications, and PacBio-BAM, a person of ordinary skill in the art would have been motivated to combine the methods taught by these references. For example, a person of ordinary skill in the art would expect that doing so would be able to provide the benefits that Tsai discusses, such as efficiently targeting regions associated with disorders to analyze their methylation status. A person of ordinary skill in the art would also have a reasonable expectation of success in such a combination, as each reference relates to the same DNA sequencing technology, combined according to known methods, and yielding routine, predictable results.

A. Claim 13

201. Claim 13 depends from claim 12 and further recites that “the sample nucleic acid molecule is a circular DNA molecule formed by cutting a double-stranded DNA molecule using a Cas9 complex to form a cut double-stranded DNA molecule, and ligating a hairpin adaptor onto an end of the cut double-stranded DNA molecule.” These elements are taught by the Flusberg Publications and Tsai.

202. Tsai teaches exposing “guide RNAs to Cas9 endonucleases ... to form a sgRNA-Cas9 complex,” binding the Cas9 complex to a nucleic acid sample, and

“subjecting the nucleic acid sample to Cas9 cleavage, wherein the Cas9 endonucleases cleave the nucleic acid sample.” EX1020, ¶7. Tsai further teaches “linking stem-loop adapters to all double-stranded ends in the nucleic acid sample after the Cas9 cleavage to form adapter-ligated fragments.” *Id.*; *see also id.*, ¶55. The result is to form a “single-stranded circular construct,” i.e., a circular DNA molecule. EX1020, ¶41. Fig. 8 of Tsai (below) illustrates this process:

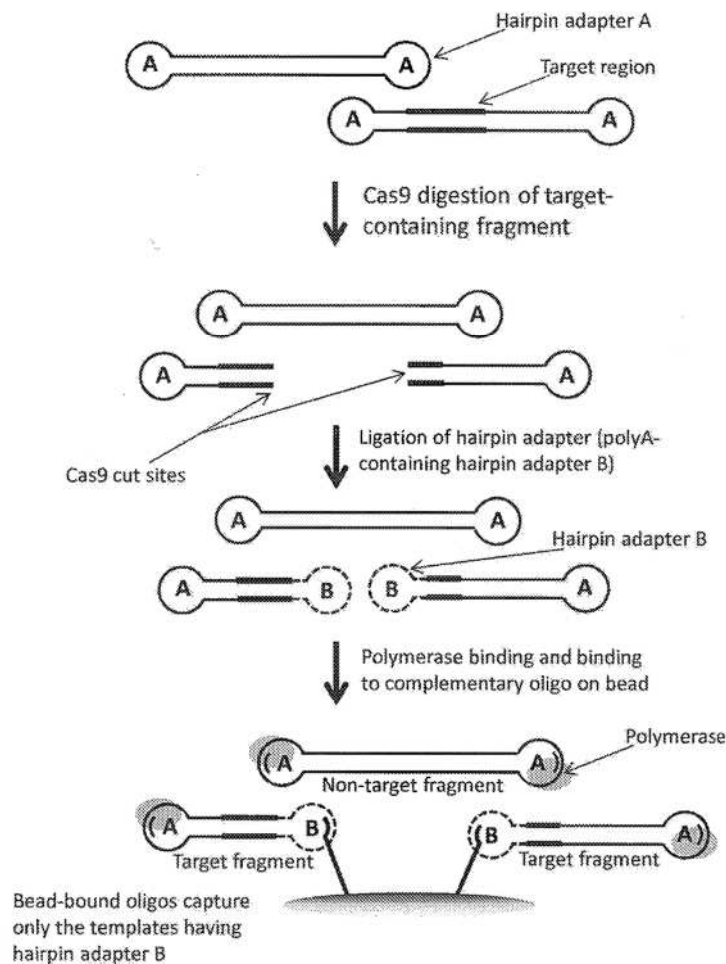


Figure 8

EX1020, Fig. 8.

'794 patent	Prior Art
13. The method of claim 12, wherein the sample nucleic acid molecule is a circular DNA molecule formed by: cutting a double-stranded DNA molecule using a Cas9 complex to form a cut double-stranded DNA molecule, and ligating a hairpin adaptor onto an end of the cut double-stranded DNA molecule.	“For example, the amplified fragments will be ligated to a first pair of hairpins, and the unamplified fragments will be ligated to a second pair of hairpins, wherein the sequence of the first pair is distinct from the sequence of the second pair, e.g., in either one or both hairpins within each pair.” EX1007, 60:54-55. “c) exposing the guide RNAs to Cas9 endonucleases such that each of the guide RNAs is bound to one of the Cas9 endonucleases to form a sgRNA-Cas9 complex; d) combining the sgRNA-Cas9 complex with the nucleic acid sample under conditions that promote binding of the sgRNA-Cas9 complex to the nucleic acid sample at the first location and the second location; e) subjecting the nucleic acid sample to Cas9 cleavage, wherein the Cas9 endonucleases cleave the nucleic acid sample at the first location and the second location; f) linking stem-loop adapters to all double-stranded ends in the nucleic acid sample after the Cas9 cleavage to form adapter-ligated fragments.” EX1020, ¶7. “[T]he nucleic acids in the mixture are linked to stem-loop (a.k.a., ‘hairpin’) adapters at one or both ends to lock the two strands together during the enrichment procedure.” EX1020, ¶55. “Further, the ligation of hairpin adapters onto both ends of the target region can provide a sequencing template that comprises both the complementary and non-complementary strands of the target region in a single-stranded circular construct...” EX1020, ¶41. <i>See also</i> EX1006, 463; EX1020, ¶¶6, 23, 65-66, 102, Fig. 8.

203. Accordingly, it is my opinion that claim 13 would have been obvious over the Flusberg Publications in view of PacBio-BAM and the Tsai.

B. Claims 15 and 16

204. As I discussed above, claim 15 depends from claim 1 and further recites “wherein each nucleotide within the window is enriched or filtered,” and claim 16 depends from claim 15 and further characterizes how the nucleotides are either enriched or filtered. The Flusberg Publications in view of PacBio-BAM and Tsai disclose or teach each element of claims 15 and 16.

205. Whereas in Section XIII.B, I focused on the “filtered” alternative for these claims, in this section I address the “enriched” alternative. For this alternative, claim 16 specifies that the nucleotides are enriched by performing the exact same steps recited in claim 13.

'794 patent	Prior Art
15. The method of claim 1, wherein each nucleotide within the window is enriched or filtered. 16. The method of claim 15, wherein each nucleotide within the window is enriched by: cutting a double-stranded DNA molecule using a Cas9 complex to form a cut double-stranded DNA molecule, and ligating a hairpin adaptor onto an end of	<i>See supra</i> Section XIV.A (discussing claim 13); EX1006, 463; EX1007, 60:54-55; EX1020, ¶¶6-7, 23, 41, 55, 65-66, 102, Fig. 8.

the cut double-stranded DNA molecule, or filtered by: selecting double- stranded DNA molecules having a size with a size range.	
---	--

206. Accordingly, it is my opinion that claims 15 and 16 would have been obvious over the Flusberg Publications in view of PacBio-BAM and the Tsai.

207. Accordingly, it is my opinion that the prior art renders claims 1-19 obvious.

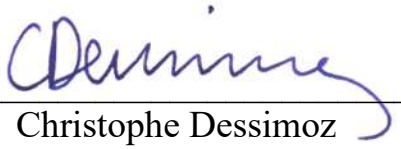
XV. CONCLUDING STATEMENTS

208. In signing this declaration, I understand that the declaration will be filed as evidence in a contested case before the Patent Trial and Appeal Board of the United States Patent and Trademark Office. I acknowledge that I may be subject to cross-examination in this case and that cross-examination will take place within the United States. If cross-examination is required of me, I will appear for cross-examination within the United States during the time allotted for cross-examination.

209. I declare that all statements made herein of my knowledge are true, and that all statements made on information and belief are believed to be true, and 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
Section 1001 of Title 18 of the United States Code.

Dated: October 17, 2023

By: 
Christophe Dessimoz

XVI. APPENDIX – LISTING OF CHALLENGED CLAIMS

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

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

for each nucleotide:

an identity of the nucleotide,

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

a width of the pulse corresponding to the nucleotide, and

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

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

the identity of the nucleotide,

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

the width of the pulse corresponding to the nucleotide, and

the interpulse duration;

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

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

2. The method of claim 1, wherein:

the input data structure is one input data structure of a plurality of input data structures,

the sample nucleic acid molecule is one sample nucleic acid molecule of a plurality of sample nucleic acid molecules,

the plurality of sample nucleic acid molecules are obtained from a biological sample of a subject, and

each input data structure corresponds to a respective window of nucleotides sequenced in a respective sample nucleic acid molecule of the plurality of sample nucleic acid molecules, and

the method further comprising:

receiving the plurality of input data structures,

inputting the plurality of input data structures into the model, and

determining, using the model, whether a modification is present in a nucleotide at a target location in the respective window of each input data structure.

3. The method of claim 2, further comprising: determining the modification is present at one or more nucleotides, and determining a classification of a disorder using the presence of the modification at one or more nucleotides.
4. The method of claim 3, wherein the disorder comprises cancer.
5. The method of claim 4, further comprising: determining that the classification of the disorder is that the subject has the disorder, and treating the subject for the disorder by chemotherapy, radiation, or surgery.
6. The method of claim 3, wherein determining the classification of the disorder uses the number of modifications or the sites of the modifications.
7. The method of claim 2, wherein the modification is a methylation, the method further comprising: determining the modification is present at one or more nucleotides, and determining a clinically-relevant DNA fraction, a fetal methylation profile, a maternal methylation profile, a presence of an imprinting gene region, or a tissue of origin using the presence of the modification at one or more nucleotides.
8. The method of claim 2, wherein each sample nucleic acid molecule of the plurality of sample nucleic acid molecules has a size greater than a cutoff size.
9. The method of claim 2, wherein: the plurality of sample nucleic acid molecules align to a plurality of genomic regions, for each genomic region of the plurality of genomic regions: a number of sample nucleic acid molecules is aligned to the genomic region, the number of sample nucleic acid molecules is greater than a cutoff number.
10. The method of claim 1, further comprising sequencing the sample nucleic acid molecule.
11. The method of claim 1, wherein the model includes a machine learning model, a principal component analysis, a convolutional neural network, or a logistic regression.
12. The method of claim 1, wherein: the window of nucleotides corresponding to the input data structure comprises nucleotides on a first strand of the sample

nucleic acid molecule and nucleotides on a second strand of the sample nucleic acid molecule, and the input data structure further comprises for each nucleotide within the window a value of a strand property, the strand property indicating the nucleotide being present on either the first strand or the second strand.

13. The method of claim 12, wherein the sample nucleic acid molecule is a circular DNA molecule formed by: cutting a double-stranded DNA molecule using a Cas9 complex to form a cut double-stranded DNA molecule, and ligating a hairpin adaptor onto an end of the cut double-stranded DNA molecule.

14. The method of claim 1, wherein the nucleotides within the window are determined using a circular consensus sequence and without alignment of the sequenced nucleotides to a reference genome.

15. The method of claim 1, wherein each nucleotide within the window is enriched or filtered.

16. The method of claim 15, wherein each nucleotide within the window is enriched by: cutting a double-stranded DNA molecule using a Cas9 complex to form a cut double-stranded DNA molecule, and ligating a hairpin adaptor onto an end of the cut double-stranded DNA molecule, or filtered by: selecting double-stranded DNA molecules having a size with a size range.

17. The method of claim 1, wherein nucleotides within the window are determined without using a circular consensus sequence and without alignment of the sequenced nucleotides to a reference genome.

18. The method of claim 1, wherein the optical signal is a fluorescence signal from a dye-labeled nucleotide.

19. The method of claim 1, wherein each window associated with the first plurality of data structures comprises 4 consecutive nucleotides on a first strand of each first nucleic acid molecule.

XVII. EXHIBIT LIST

Exhibit No.	Description
1001	U.S. Patent No. 11,091,794 to Lo (“the ’794 patent”)
1003	Curriculum Vitae of Christophe Dessimoz, Ph.D.
1006	Flusberg, et al., “Direct Detection of DNA Methylation During Single-molecule, Real-Time Sequencing,” <i>Nature Methods</i> 7(6) (2010) (“Flusberg”)
1007	U.S. Patent No. 9,175,338 to Flusberg et al. (“Flusberg-338”)
1008	Clark, et al. “Direct Detection and Sequencing of Damaged DNA Bases” <i>Genome Integrity</i> 2 (2011) (“Clark”)
1009	Loomis et al., “Sequencing the Unsequenceable: Expanded CGG-repeat Alleles of the Fragile X Gene,” <i>Genome Research</i> 23(1) (2013) (“Loomis”)
1010	Schadt, et al., “Modeling Kinetic Rate Variation in Third Generation DNA Sequencing Data to Detect Putative Modifications to DNA Bases,” <i>Genome Research</i> 23(1) (2013) (“Schadt”)
1011	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 Sep. 16, 2017) (“PacBio-BAM”)
1012	SAM/BAM Specification Working Group, Sequence Alignment/Map Format Specification, http://web.archive.org/web/20180613134937/http://samtools.github.io/hts-specs/SAMv1.pdf (archived Jun. 13, 2018) (“SAM/BAM Specification”)
1013	Davis et al., “Entering the Era of Bacterial Epigenomics with Single Molecule Real Time DNA Sequencing,” <i>Current Opinion in Microbiology</i> , 16(2) (2013) (“Davis”)

1014	<i>Take2 Tech, Ltd. v. Pac. Biosciences of Cal., Inc.</i> , DDE-1-22-cv-01595, Infringement Claim Chart for U.S. Patent No. 11,091,794 (“Take2 Infringement Contentions”)
1015	Feng, et al, “Detecting DNA modifications from SMRT Sequencing Data by Modeling Sequence Context Dependence of Polymerase Kinetic,” <i>PLoS Computational Biology</i> , 9(3) (2013) (“Feng”)
1016	“Detecting DNA Base Modifications Using Single Molecule, Real-Time Sequencing,” https://www.pacb.com/wp-content/uploads/2015/09/WP_Detecting_DNA_Base_Modifications_Using_SMRT_Sequencing.pdf (2012) (“PacBio White Paper”)
1017	Zou et al., “A Primer on Deep Learning in Genomics” <i>Nature Genetics</i> , 51(1) (Jan. 2019)
1018	U.S. Patent Pub. No. 2016/0326593 to Clement et al. (“Clement”)
1019	Pacific Biosciences Template Preparation and Sequencing Guide, https://www.pacb.com/wp-content/uploads/2015/09/Guide-Pacific-Biosciences-Template-Preparation-and-Sequencing.pdf (2015) (“PacBio Template Guide”)
1020	U.S. Patent Publication No. 2016/0208241 to Tsai et al. (“Tsai”)
1021	Pacific Biosciences kineticsTools Manual, http://web.archive.org/web/20170813085437/https://github.com/PacificBiosciences/kineticsTools/blob/master/doc/manual.rst (2018) (“kineticsTools manual”)
1022	Liu et al., “Detection of DNA Base Modifications by Deep Recurrent Neural Network on Oxford Nanopore Sequencing Data,” <i>Nature Comms.</i> 10(1) (Jun. 4, 2019) (“Q. Liu”)
1023	Eid et al., “Real-time DNA sequencing from single polymerase molecules,” <i>Science</i> 323(5910) (2009)
1024	Int. Appl. Publ. No. WO 2012/065043 (“Flusberg PCT”)
1025	Jonas Korlach & Stephen W. Turner, “Single-Molecule Sequencing” in <i>Encyclopedia of Biophysics</i> (2013) (“Korlach”)

1026	GenomeWeb, “PacBio Reveals Beta System Specs for RS; Says Commercial Release is on Track for First Half of 2011,” https://www.genomeweb.com/sequencing/pacbios-q3-losses-widen-it-readies-launch-rs-system (2010) (“GenomeWeb”)
1027	Jacob Schreiber et al., “Error Rates for Nanopore Discrimination among Cytosine, Methylcytosine, and Hydroxymethylcytosine along Individual DNA Strands,” <i>Proceedings of the National Academy of Sciences</i> 110(47) (2013) (“Schreiber”)
1028	Langlotz, A Roadmap for Foundational Research on Artificial Intelligence in Medical Imaging: From the 2018 NIH/RSNA/ACR/The Academy Workshop, <i>Radiology</i> 291(3) (Apr. 16, 2019) (“Langlotz”)
1029	Miotto et al., “Deep Learning for Healthcare: Review, Opportunities and Challenges,” <i>Briefings in Bioinformatics</i> , 19(6) (2018) (“Miotto”)
1030	Tse et al., “Genome-wide Detection of Cytosine Methylation by Single Molecule Real-time Sequencing,” <i>Proceedings of the National Academy of Sciences</i> , 118(5) (2021) (“Tse”)
1031	Tiernan Ray, “AI is Changing the Entire Nature of Compute,” <i>ZDNet</i> , http://web.archive.org/web/20190630155704/https://www.zdnet.com/article/ai-is-changing-the-entire-nature-of-compute/ (June 30, 2019) (“Ray”)
1032	Wenger et al., “Accurate Circular Consensus Long-read Sequencing Improves Variant Detection and Assembly of a Human Genome,” <i>Nature Biotechnology</i> , 37(10) (Aug. 12, 2019) (“Wenger”)
1033	Pacific Biosciences, PacBioFileFormats Documentation (Feb. 2022) (“PacBioFileFormats 2022”)
1034	Lau et al., “LongISLND: <i>in silico</i> Sequencing of Lengthy and Noisy Datatypes,” <i>Bioinformatics</i> , 32(24) (2016) (“Lau”)
1035	Tahir et al., “iRNA-PseKNC(2methyl): Identify RNA 2’-O-methylation Sites by Convolution Neural Network and Chou’s Pseudo Components,” <i>Journal of Theoretical Biology</i> , 465 (2018) (“Tahir”)

1036	DeepVariant Usage Guide, https://github.com/google/deepvariant/blob/42b9c777436d8306f6cb9d2e18da756fa370fcd8/docs/deepvariant-details.md (“DeepVariant”)
1038	Liu et al., “Accurate Detection of m6A RNA Modifications in Native RNA Sequences,” https://www.biorxiv.org/content/10.1101/525741v1 (Jan. 21, 2019) (“H. Liu”)
1042	GitHub, PacBio BAM 3.0, https://github.com/PacificBiosciences/PacBioFileFormats/blob/3.0/BAM.rst (Mar. 28, 2016)
1043	GitHub, PacBio BAM 5.0, https://github.com/PacificBiosciences/PacBioFileFormats/blob/5.0/BAM.rst (Mar. 7, 2017)
1044	PacBio Sequencing – How it Works