

BRIEF SUMMARY

[0007] We have developed a new method that, in one example, allows the determination of base modifications, such as 5mC in nucleic acids without template DNA pre-treatment such as enzymatic and/or chemical conversions, or protein and/or antibody binding. While such template DNA pre-treatment is not necessary for the determination of the base modifications, in examples that are shown, certain pre-treatment (e.g. digestion with restriction enzymes) may serve to enhance examples of the disclosure (e.g. allowing the enrichment of CpG sites for analysis). The examples present in this disclosure could be used for detecting different types of base modification, for example, including but not limited to 4mC, 5hmC, 5fC, and 5caC, 1mA, 3mA, 7mA, 3mC, 2mG, 6mG, 7mG, 3mT, and 4mT, etc. Such examples can make use of features derived from sequencing, such as kinetic features, that are affected by the various base modifications, as well as an identity of nucleotides in a window around a target position whose methylation status is determined.

[0008] Examples of the present disclosure can be used for, but is not limited to, single molecule sequencing. One type of single molecule sequencing is single molecule, real-time sequencing in which the progress of the sequencing of a single DNA molecule is monitored in real-time. One type of single molecule, real-time sequencing is that commercialized by Pacific Biosciences using their Single Molecule, Real-Time (SMRT) system. Methods may use the pulse width of a signal from sequencing bases, the interpulse duration (IPD) of bases, and the identity of the bases in order to detect a modification in a base or in a neighboring base. Another single molecule system is that based on nanopore sequencing. One example of a nanopore sequencing system is that commercialized by Oxford Nanopore Technologies.

[0009] The methods we have developed can serve as tools to detect base modifications in biological samples to assess the methylation profiles in the samples for various purposes including but not limited to research and diagnostic purposes. The detected methylation profiles can be used for different analysis. The methylation profiles can be used to detect the origin of DNA (e.g., maternal or fetal, tissue, bacterial, or DNA obtained from tumor cells enriched from the blood of a cancer patient). Detection of aberrant methylation profiles in tissues aids the

identification of developmental disorders in individuals, identify and prognosticate tumors or malignancies.

[0010] Examples of the present disclosure may include analyzing the relative methylation levels of haplotypes of an organism. An imbalance in the methylation levels between the two haplotypes may be used to determine a classification of a disorder. A higher imbalance may indicate the presence of a disorder or a more severe disorder. The disorder may include cancer.

[0011] Methylation patterns in a single molecule can identify chimera and hybrid DNA. Chimeric and hybrid molecules may include sequences from two different genes, chromosomes, organelles (e.g. mitochondria, nucleus, chloroplasts), organisms (mammals, bacteria, viruses, etc.), and/or species. Detecting junctions of chimeric or hybrid DNA molecules may allow for detecting gene fusions for various disorders or diseases, including cancer, prenatal, or congenital disorders.

[0012] A better understanding of the nature and advantages of examples of the present disclosure may be gained with reference to the following detailed description and the accompanying drawings.

[0012a] In one aspect, disclosed herein is a method for detection of a methylation of a nucleotide in a nucleic acid molecule, the method comprising: 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 the respective window comprises at least two nucleotides, wherein each of the first nucleic acid molecules is sequenced by measuring pulses in an optical signal corresponding to the nucleotides, wherein the methylation has a known first state in the nucleotide at a target position in each window of each first nucleic acid molecule, each first data structure comprising values for the following properties: for each nucleotide within the window: an identity of the nucleotide, a position of the nucleotide with respect to the target position within the respective window, 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, 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 for the methylation of the nucleotide at the target position; and training a model using the plurality of first training samples by optimizing 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 methylation.

[0012b] In another aspect, disclosed herein is a computer product comprising a non-transitory computer readable medium storing a plurality of instructions that when executed control a computer system to perform the method of the disclosure.

[0012c] In another aspect, disclosed herein is a system comprising: the computer product of the disclosure; and one or more processors for executing instructions stored on the computer readable medium.

[0012d] In another aspect, disclosed herein is a method for detecting a methylation of a nucleotide in a nucleic acid molecule, the method comprising: receiving input data acquired by measuring pulses in an optical signal corresponding to a window of nucleotides in a sample nucleic acid molecule, and wherein the window comprises at least two nucleotides, the input data comprising values for the following properties: for each nucleotide within the window: an identity of the nucleotide, a position of the nucleotide with respect to a target position within the respective window, 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; determining whether the methylation is present in a nucleotide at the target position within the window using: for each nucleotide within the window: the identity of the nucleotide, the position of the nucleotide with respect to the target position, the width of the pulse corresponding to the nucleotide, and the interpulse duration.

[0012e] In another aspect, disclosed herein is a computer product comprising a non-transitory computer readable medium storing a plurality of instructions that when executed control a computer system to perform the method of the disclosure.

[0012f] In another aspect, disclosed herein is a system comprising: the computer product of the disclosure; and one or more processors for executing instructions stored on the computer readable medium.

5

BRIEF DESCRIPTION OF THE DRAWINGS

[0013] **FIG. 1** illustrates SMRT sequencing of molecules carrying base modifications according to embodiments of the present invention.

[0014] **FIG. 2** illustrates SMRT sequencing of molecules carrying methylated and unmethylated CpG sites according to embodiments of the present invention.

10

[0015] **FIG. 3** illustrates interpulse durations and pulse width according to embodiments of the present invention.

[0016] **FIG. 4** shows an example of a measurement window of the Watson strand of DNA for detecting a base modification according to embodiments of the present invention.

WHAT IS CLAIMED IS:

- 1 1. A method for detection of a methylation of a nucleotide in a nucleic acid
2 molecule, the method comprising:
3 receiving a first plurality of first data structures, each first data structure of
4 the first plurality of first data structures corresponding to a respective window of
5 nucleotides sequenced in a respective nucleic acid molecule of a plurality of first nucleic
6 acid molecules, wherein the respective window comprises at least two nucleotides,
7 wherein each of the first nucleic acid molecules is sequenced by measuring pulses in an
8 optical signal corresponding to the nucleotides, wherein the methylation has a known first
9 state in the nucleotide at a target position in each window of each first nucleic acid
10 molecule, each first data structure comprising values for the following properties:
11 for each nucleotide within the window:
12 an identity of the nucleotide,
13 a position of the nucleotide with respect to the target position within the
14 respective window,
15 a width of the pulse corresponding to the nucleotide, and
16 an interpulse duration representing a time between the pulse
17 corresponding to the nucleotide and a pulse corresponding to a neighboring nucleotide,
18 storing a plurality of first training samples, each including one of the first
19 plurality of first data structures and a first label indicating the first state for the
20 methylation of the nucleotide at the target position; and
21
22 training a model using the plurality of first training samples by optimizing
23 parameters of the model based on outputs of the model matching or not matching corresponding
24 labels of the first labels when the first plurality of first data structures is input to the model,
25 wherein an output of the model specifies whether the nucleotide at the target position in the
26 respective window has the methylation.
- 1 2. The method of claim 1, further comprising:
2 receiving a second plurality of second data structures, each second data structure
3 of the second plurality of second data structures corresponding to a respective window of

nucleotides sequenced in a respective nucleic acid molecule of a plurality of second nucleic acid molecules, wherein the methylation has a known second state in a nucleotide at a target position within each window of each second nucleic acid molecule, each second data structure comprising values for the same properties as the first plurality of first data structures;

storing a plurality of second training samples, each including one of the second plurality of second data structures and a second label indicating the second state of the nucleotide at the target position;

wherein:

the first state or the second state is that the methylation is present and the other state is that the methylation is absent, and training the model further comprises using the plurality of second training samples by optimizing parameters of the model based on outputs of the model matching or not matching corresponding labels of the second labels when the second plurality of second data structures are input to the model.

3. The method of claim 2, wherein the plurality of first nucleic acid molecules is the same as the plurality of the second nucleic acid molecules.

4. The method of claim 2, wherein the plurality of first nucleic acid molecules is generated using multiple displacement amplification with methylated nucleotides of a first type, and wherein the plurality of second nucleic acid molecules is generated using multiple displacement amplification with unmethylated nucleotides of the first type.

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

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

7. The method of claim 6, wherein the windows associated with the first plurality of first data structures comprise the same number of consecutive nucleotides.

8. The method of claim 1, wherein:

each window associated with the first plurality of data structures comprises nucleotides on a first strand of the first nucleic acid molecule and nucleotides on a second strand of the first nucleic acid molecule, and
each first 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.

9. The method of claim 1, wherein the neighboring nucleotide is the adjacent nucleotide.

10. The method of claim 1, wherein the width of the pulse is the width of the pulse at half the maximum value of the pulse.

11. The method of claim 1, wherein the interpulse duration is the time between the maximum value of the pulse associated with the nucleotide and the maximum value of the pulse associated with the neighboring nucleotide.

12. The method of claim 1, wherein the model comprises a convolutional neural network comprising:

a set of convolutional filters configured to filter the first plurality of first data structures,

an input layer configured to receive the filtered first plurality of first data structures,

a plurality of hidden layers including a plurality of nodes, a first layer of the plurality of hidden layers coupled to the input layer; and

an output layer coupled to a last layer of the plurality of hidden layers and configured to output an output data structure, the output data structure comprising the properties.

13. The method of claim 1, wherein the known first states include a methylated state for a first portion of the first data structures and an unmethylated state for a second portion of the first data structures.

14. The method of claim 1, wherein the methylation comprises 4mC (N4-methylcytosine), 5mC (5-methylcytosine), 5hmC (5-hydroxymethylcytosine), 5fC (5-

3 formylcytosine), 5caC (5-carboxylcytosine), 1mA (N1-methyladenine), 3mA (N3-
4 methyladenine), 6mA (N6-methyladenine), 7mA (N7-methyladenine), 3mC (N3-
5 methylcytosine), 2mG (N2-methylguanine), 6mG (O6-methylguanine), 7mG (N7-
6 methylguanine), 3mT (N3-methylthymine), or 4mT (O4-methylthymine).

15. The method of claim 1, wherein the methylation comprises 5mC (5-methylcytosine) or 6mA (N6-methyladenine).

16. The method of claim 1, wherein the methylation comprises 5mC (5-methylcytosine).

1 17. The method of claim 1, wherein each data structure further comprises a
2 value for a height of the pulse corresponding to each nucleotide within the window.

1 18. The method of claim 1, wherein the optical signal corresponding to the
2 nucleotides results from the nucleotides or a tag associated with the nucleotides.

1 19. The method of claim 18, wherein each target position is the center of the
2 respective window.

1 20. The method of claim 1, wherein the methylation is absent in each window
2 of each first nucleic acid molecule.

1 21. The method of claim 1, wherein:
2 each first data structure of the first plurality of first data structures excludes first
3 nucleic acid molecules with an interpulse duration or a width of a pulse below a cutoff value.

1 22. The method of claim 1, wherein
2 the plurality of first training samples is generated by:
3 amplifying a plurality of nucleic acid molecules using a set of nucleotides,
4 wherein the set of nucleotides includes 6mA at a specified ratio.

1 23. The method of claim 22, wherein the methylation comprises 6mA (N6-
2 methyladenine).

1 24. The method of claim 1, wherein at least some of the plurality of first
2 nucleic acid molecules each include a first portion corresponding to a first reference sequence
3 and a second portion corresponding to a second reference sequence that is disjoint from the first
4 reference sequence.

1 25. The method of claim 1, further comprising:
2 validating the model using a plurality of chimeric nucleic acid molecules, each
3 including a first portion corresponding to a first reference sequence and a second portion
4 corresponding to a second reference sequence, wherein the first portion has a first methylation
5 pattern and the second portion has a second methylation pattern.

1 26. The method of claim 24 or claim 25, wherein the first portion is treated
2 with a methylase.

1 27. The method of claim 26, wherein the second portion corresponds to an
2 unmethylated portion of the second reference sequence.

1 28. The method of claim 24 or claim 25, wherein the first reference sequence
2 is human, and wherein the second reference sequence is from a different animal.

1 29. A computer product comprising a non-transitory computer readable
2 medium storing a plurality of instructions that when executed control a computer system to
3 perform the method of any one of the preceding claims.

1 30. A system comprising:
2 the computer product of claim 29; and
3 one or more processors for executing instructions stored on the computer readable
4 medium.

 31. A method for detecting a methylation of a nucleotide in a nucleic acid
molecule, the method comprising:
 receiving input data acquired by measuring pulses in an optical signal
corresponding to a window of nucleotides in a sample nucleic acid molecule, and wherein the

window comprises at least two nucleotides, the input data comprising values for the following properties:

for each nucleotide within the window:

an identity of the nucleotide,

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

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;

determining whether the methylation is present in a nucleotide at the target position within the window using:

for each nucleotide within the window:

the identity of the nucleotide,

the position of the nucleotide with respect to the target position,

the width of the pulse corresponding to the nucleotide, and

the interpulse duration.

32. The method of claim 31, wherein determining whether the methylation is present in the nucleotide at the target position within the window comprises inputting the input data into a model, the model configured to:

receive the input data,

apply optimized parameters of the model to the properties of the input data; and

output a methylation status specifying whether the nucleotide at the target position in the respective window has a methylation.

1 33. The method of claim 31, wherein determining whether the methylation is
2 present in the nucleotide at the target position within the window comprises inputting the input
3 data into a model, the model trained by:

4 receiving additional data, the additional data acquired by measuring pulses in
5 optical signals corresponding to a plurality of windows of nucleotides, each window of
6 nucleotides corresponding to a first nucleic acid molecule of a plurality of first nucleic acid
7 molecules, wherein the methylation has a known first state in a nucleotide at a target position in

8 each window of each first nucleic acid molecule, the additional data comprising values for the
9 properties,

10 storing a plurality of first training samples, each including values for the
11 properties of each first nucleic acid molecule and a first label indicating the first state of the
12 nucleotide at the target position, and

13 optimizing, using the plurality of first training samples, parameters of the model
14 based on outputs of the model matching or not matching corresponding labels of the first labels
15 when the values of the properties of the first training samples are input to the model, wherein an
16 output of the model specifies whether the nucleotide at the target position in the respective
17 window has the methylation.

34. The method of claim 32 or 33, wherein the model includes a machine learning model, a principal component analysis, a convolutional neural network, or a logistic regression.

35. The method of any one of claims 31 to 34, wherein the values for the properties are arranged in a matrix structure.

36. The method of any one of claims 31 to 35, wherein:

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 is obtained from a biological sample of a subject, and

the input data comprises data acquired by measuring pulses in optical signals corresponding to a respective window of nucleotides in each sample nucleic acid molecule of the plurality of sample nucleic acid molecules, and

the method further comprising:

for each sample nucleic acid of the plurality of sample nucleic acid molecules, determining whether a methylation is present in a nucleotide at a respective target location in the respective window using:

for each nucleotide within the respective window:

the identity of the nucleotide,
the position of the nucleotide with respect to the respective target position,
the width of the pulse corresponding to the nucleotide, and
the interpulse duration.

37. The method of claim 36, further comprising:
determining the methylation is present at one or more nucleotides, and
determining a classification of a disorder in a subject using the presence of the
methylation at one or more nucleotides, wherein the sample nucleic acid molecule is obtained
from the subject.

38. The method of claim 37, wherein the disorder comprises cancer.

39. The method of claim 37 or 38, wherein determining the classification of
the disorder uses the number of methylations or the sites of the methylations.

40. The method of claim 36, further comprising:
determining a methylation status for whether the methylation 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 methylation status at the one or more nucleotides.

41. The method of claim 40, wherein:
the method comprises determining the tissue of origin, and
determining the tissue of origin comprises determining whether the sample
nucleic acid molecule is of fetal or maternal origin.

42. The method of claim 41, wherein determining whether the sample nucleic
acid molecule is of fetal or maternal origin comprises:
determining a methylation level of the sample nucleic acid molecule using the
methylation statuses of the one or more nucleotides, and
comparing the methylation level of the sample nucleic acid molecule to a
reference value.

43. The method of any one of claims 36 to 42, further comprising filtering the
biological sample for nucleic acid molecules having sizes greater than a cutoff size to obtain the
plurality of sample nucleic acid molecules.

44. The method of any one of claims 36 to 43, further comprising:
aligning the plurality of sample nucleic acid molecules to a plurality of genomic regions, wherein each genomic region of the plurality of genomic regions has a subread depth greater than a cutoff number.

45. The method of any one of claims 31 to 44, further comprising sequencing the sample nucleic acid molecule.

46. The method of any one of claims 31 to 45, wherein:
the window of nucleotides 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 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.

47. The method of claim 46, 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.

48. The method of claim 31, wherein the nucleotides within the window are determined using a circular consensus sequence and are determined without alignment of the sequenced nucleotides to a reference genome.

49. The method of claim 31, wherein each nucleotide within the window is enriched or filtered.

50. The method of claim 49, 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.

51. The method of claim 31, wherein nucleotides within the window are determined without using a circular consensus sequence and without alignment of the sequenced nucleotides to a reference genome.

52. The method of any one of claims 31 to 51, further comprising using the presence of the methylation to identify a disorder in a subject, to identify or prognosticate tumors or malignancies in the subject, to detect the tissue origin of the sample nucleic acid molecule, or to identify chimera and hybrid DNA, wherein the sample nucleic acid molecule is obtained from the subject.

53. A computer product comprising a non-transitory computer readable medium storing a plurality of instructions that when executed control a computer system to perform the method of any one of claims 31 to 52.

1 54. A system comprising:
2 the computer product of claim 53; and
3 one or more processors for executing instructions stored on the computer readable
4 medium

1