

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.

[0007a] In one aspect, the present disclosure refers to a method for detecting a methylation 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 window comprises at least two nucleotides, and 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 methylation 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 methylation, and (d) determining, using the model, whether the methylation is present in a nucleotide at the target position within the window in the input data structure.

[0007b] In another aspect, the present disclosure refers to 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 as disclosed herein.

[0007c] In another aspect, the present disclosure refers to a system comprising: the computer product as disclosed herein; and one or more processors for executing instructions stored on the computer readable medium.

[0007d] In another aspect, the present disclosure refers to a system comprising means for performing the method as disclosed herein.

[0007e] In another aspect, the present disclosure refers to a system comprising one or more processors configured to perform the method as disclosed herein.

[0007f] In another aspect, the present disclosure refers to a system comprising modules that respectively perform the steps of the method as disclosed herein.

[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

CLAIMS:

1. A method for detecting a methylation 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 window comprises at least two nucleotides, and 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 methylation 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 methylation, and

(d) determining, using the model, whether the methylation 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 methylation 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 methylation is present at one or more nucleotides, and

determining a classification of a disorder using the presence of the methylation at one or more nucleotides.

4. The method of claim 3, wherein the disorder comprises cancer.

5. The method of claim 3 or 4, further comprising:
determining that the classification of the disorder is that the subject has the disorder.

6. The method of any one of claims 3 to 5, wherein determining the classification of the disorder uses the number of methylations or the sites of the methylations.

7. The method of claim 2, 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.

8. The method of claim 7, 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.

9. The method of claim 8, 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.

10. The method of any one of claims 2 to 9, wherein each sample nucleic acid molecule of the plurality of sample nucleic acid molecules has a size greater than a cutoff size.

11. The method of any one of claims 2 to 10, 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.
12. The method of any one of claims 1 to 11, further comprising sequencing the sample nucleic acid molecule.
13. The method of any one of claims 1 to 12, wherein the model includes a machine learning model, a principal component analysis, a convolutional neural network, or a logistic regression.
14. The method of any one of claims 1 to 13, 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.
15. The method of claim 14, 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.
16. The method of any one of claims 1 to 15, wherein the nucleotides within the window are determined using a circular consensus sequence and without alignment of the sequenced nucleotides to a reference genome.

17. method of any one of claims 1 to 16, wherein each nucleotide within the window is enriched or filtered.

18. The method of claim 17, 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.

19. The method of any one of claims 1 to 13, wherein nucleotides within the window are determined without using a circular consensus sequence and without alignment of the sequenced nucleotides to a reference genome.

20. The method of any one of claims 1 to 19, wherein the optical signal is a fluorescence signal from a dye-labeled nucleotide.

21. The method of any one of claim 1 to 20, 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.

22. The method of any one of claims 1 to 21, wherein the methylation comprises 4mC (N4-methylcytosine), 5mC (5-methylcytosine), 5hmC (5-hydroxymethylcytosine), 5fC (5-formylcytosine), 5caC (5-carboxylcytosine), 1mA (N1-methyladenine), 3mA (N3-methyladenine), 6mA (N6-methyladenine), 7mA (N7-methyladenine), 3mC (N3-methylcytosine), 2mG (N2-methylguanine), 6mG (O6-methylguanine), 7mG (N7-methylguanine), 3mT (N3-methylthymine), or 4mT (O4-methylthymine).

23. The method of any one of claims 1 to 21, wherein the methylation comprises 5mC (5-methylcytosine).

24. The method of any one of claims 1 to 21, wherein the methylation comprises 6mA (N6-methyladenine).

25. The method of claim 1, 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 is obtained from the subject.

26. The method of any one of claims 1 to 25, wherein at least some of the plurality of first nucleic acid molecules each include a first portion corresponding to a first reference sequence and a second portion corresponding to a second reference sequence that is disjoint from the first reference sequence.

27. The method of any one of claims 1 to 26, further comprising:
validating the model using a plurality of chimeric nucleic acid molecules, each including a first portion corresponding to a first reference sequence and a second portion corresponding to a second reference sequence, wherein the first portion has a first methylation pattern and the second portion has a second methylation pattern.

28. The method of claim 26 or 27, wherein the first portion is treated with a methylase.

29. The method of claim 28, wherein the second portion corresponds to an unmethylated portion of the second reference sequence.

30. The method of claim 26 or 27, wherein the first reference sequence is human, and wherein the second reference sequence is from a different animal.

31. The method of any one of claims 1 to 30, wherein the window comprises at least three nucleotides downstream of a target position within the window.

32. 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 the preceding claims.

33. A system comprising:
the computer product of claim 32; and
one or more processors for executing instructions stored on the computer readable
medium.

34. A system comprising means for performing the method of any one of
claims 1-31.

35. A system comprising one or more processors configured to perform the
method of any one of claims 1-31.

36. A system comprising modules that respectively perform the steps of
the method of any one of claims 1-31.