

location in a particular viral genome. A methylation status of the viral DNA at one or more sites in the viral genome can be determined. Mixture methylation level(s) can be measured based on one or more amounts of the plurality of cell-free DNA molecules methylated at a set of site(s) of the particular viral genome. The mixture methylation level(s) can be determined in various ways, e.g., as a percentage/density of cell-free DNA molecules that are methylated at a particular site or across multiple sites, and potentially across multiple regions, each including one or more sites.

[0005] The mixture methylation level(s) can be compared to reference methylation level(s), e.g., determined from at least two cohorts of other subjects. The cohorts can have different classifications (including the first condition) associated with the particular viral genome. Other cohort(s) can correspond to other condition(s). The comparison can be performed in a variety of ways, e.g., by forming a multidimensional point of N methylation levels and determining differences from N reference methylation levels. A first classification of whether the subject has the first condition can be determined based on the comparing.

[0005a] In one aspect, the present disclosure refers to a method of analyzing a biological sample of a subject that is a human, the biological sample including a mixture of cell-free DNA molecules from a genome of the subject and from one or more other genomes, the method comprising:

analyzing a plurality of cell-free DNA molecules from the biological sample, wherein analyzing one of the plurality of cell-free DNA molecules includes:

identifying a location of the cell-free DNA molecule in a particular viral genome of a virus; and determining whether the cell-free DNA molecule is methylated at one or more sites of the particular viral genome;

measuring a mixture methylation level based on amounts of the plurality of cell-free DNA molecules methylated at a set of sites of the particular viral genome, wherein the set of sites are across a plurality of regions and are CpG sites;

comparing the mixture methylation level to one or more reference methylation levels determined from at least two cohorts of other subjects, wherein the at least two cohorts have different classifications associated with the particular viral genome, the different classifications including a first condition that is associated with the virus; and

determining a first classification of whether the subject has the first condition based on the comparing.

[0005b] In another aspect, the present disclosure refers to a computer product comprising a computer readable medium storing a plurality of instructions for controlling a computer system to perform operations of the methods as disclosed herein.

[0005c] 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.

[0005d] In another aspect, the present disclosure refers to a system comprising means for performing any of the methods as disclosed herein.

[0005e] In another aspect, the present disclosure refers to a system configured to perform any of the methods as disclosed herein.

[0005f] In another aspect, the present disclosure refers to a system comprising modules that respectively perform the steps of any of the methods as disclosed herein.

[0006] These and other embodiments of the disclosure are described in detail below. For example, other embodiments are directed to systems, devices, and computer readable media associated with methods described herein.

[0007] A better understanding of the nature and advantages of embodiments of the present disclosure may be gained with reference to the following detailed description and the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0008] FIG. 1 shows the design of capture probes for targeted bisulfite sequencing according to embodiments of the present disclosure.

[0009] FIG. 2 shows the methylation densities of CpG sites across the Epstein-Barr virus (EBV) genome in patients with infectious mononucleosis, nasopharyngeal carcinoma (NPC) and Natural killer (NK)-T cell lymphoma according to embodiments of the present disclosure.

[0010] FIG. 3 shows the methylation profiles of plasma EBV DNA in a patient (AL038) from a screening cohort with early stage NPC (Stage I) according to embodiments of the present disclosure.

WHAT IS CLAIMED IS:

1. A method of analyzing a biological sample of a subject that is a human, the biological sample including a mixture of cell-free DNA molecules from a genome of the subject and from one or more other genomes, the method comprising:

analyzing a plurality of cell-free DNA molecules from the biological sample, wherein analyzing one of the plurality of cell-free DNA molecules includes:

identifying a location of the cell-free DNA molecule in a particular viral genome of a virus; and

determining whether the cell-free DNA molecule is methylated at one or more sites of the particular viral genome;

measuring a mixture methylation level based on amounts of the plurality of cell-free DNA molecules methylated at a set of sites of the particular viral genome, wherein the set of sites are across a plurality of regions and are CpG sites;

comparing the mixture methylation level to one or more reference methylation levels determined from at least two cohorts of other subjects, wherein the at least two cohorts have different classifications associated with the particular viral genome, the different classifications including a first condition that is associated with the virus; and

determining a first classification of whether the subject has the first condition based on the comparing.

2. The method of claim 1, wherein the different classifications of the at least two cohorts further includes a second condition, the method further comprising:

determining a second classification of whether the subject has the second condition based on the comparing.

3. The method of claim 2, wherein the mixture methylation level is compared to a plurality of reference methylation levels that include a first reference methylation level and a second reference methylation level, wherein the first reference methylation level is used to determine the first classification of whether the subject has the first condition, and wherein the second reference methylation level is used to determine the second classification of whether the subject has the second condition.

4. The method of claim 3, wherein the particular viral genome is of Epstein-Barr virus, wherein the first condition is nasopharyngeal cancer, and wherein the second condition is infectious mononucleosis.

5. The method of claim 1, wherein the first classification is that the subject does not have the first condition.

6. The method of claim 1, wherein determining the first classification includes determining a level of the first condition.

7. The method of claim 1, wherein the particular viral genome is of Epstein-Barr virus, and wherein the first condition is nasopharyngeal cancer.

8. The method of claim 1, wherein the step of measuring the mixture methylation level comprises measuring N mixture methylation levels, N being an integer greater than one, wherein the mixture methylation level is one of the N mixture methylation levels, wherein each of the N mixture methylation levels is measured using a respective set of sites across a respective plurality of regions, and

wherein the step of comparing the mixture methylation level to the one or more reference methylation levels comprises comparing the N mixture methylation levels to N reference methylation levels by:

measuring differences between the N mixture methylation levels and the N reference methylation levels; and

determining whether the subject belongs to one of the at least two cohorts using the differences.

9. The method of claim 8, wherein determining whether the subject belongs to one of the at least two cohorts using the differences includes performing a hierarchical clustering analysis.

10. The method of claim 8, wherein each of the N mixture methylation levels is measured for one of a plurality of predetermined regions.

11. The method of claim 10, wherein the plurality of predetermined regions are of a same size and span the particular viral genome, and wherein the same size is between 50 bases and 1,000 bases.

12. The method of claim 10, wherein each of the plurality of predetermined regions satisfies one or more criteria including (1) a difference in methylation level among multiple subjects of a same cohort and/or (2) a difference in methylation level between a subject of one cohort and a subject of another cohort.

13. The method of claim 1, wherein each of the plurality of regions satisfies one or more criteria including (1) a difference in methylation level among multiple subjects of a same cohort and/or (2) a difference in methylation level between a subject of one cohort and a subject of another cohort.

14. The method of claim 1, wherein the set of sites satisfies one or more criteria including (1) a difference in methylation level among multiple subjects of a same cohort and/or (2) a difference in methylation level between a subject of one cohort and a subject of another cohort.

15. The method of claim 1, wherein comparing the mixture methylation level to the one or more reference methylation levels determined from at least two cohorts of other subjects comprises:

inputting the mixture methylation level into a machine learning model that was trained using the one or more reference methylation levels determined from the at least two cohorts of other subjects.

16. The method of claim 1, further comprising:

for each site of the set of sites:

determining a respective number of DNA molecules that are methylated at the site, thereby determining the amounts of the plurality of cell-free DNA molecules methylated at the set of sites of the particular viral genome.

17. The method of claim 16, further comprising:

performing methylation-aware sequencing of the plurality of cell-free DNA molecules to obtain sequence reads; and

aligning the sequence reads to the particular viral genome to determine the respective number of DNA molecules that are methylated at each site of the set of sites.

18. The method of claim 1, further comprising:

performing a methylation-aware assay of the plurality of cell-free DNA molecules as part of determining locations of the plurality of cell-free DNA molecules and whether the plurality of cell-free DNA molecules are methylated at the set of sites.

19. The method of claim 1, wherein identifying the location of the cell-free DNA molecule includes determining that the location corresponds to one of the set of sites.

20. The method of claim 1, wherein a group of the plurality of cell-free DNA molecules are analyzed collectively to determine the amounts of the plurality of cell-free DNA molecules methylated at the set of sites of the particular viral genome.

21. The method of claim 1, wherein the plurality of cell-free DNA molecules includes at least 10 cell-free DNA molecules located in the particular viral genome.

22. The method of claim 1, wherein the particular viral genome corresponds to Epstein-Barr virus, human papillomavirus, or hepatitis B virus.

23. The method of claim 1, further comprising:

for each of a set of cell-free DNA molecules in a sample:

measuring a size of the cell-free DNA molecule; and

identifying a location of the cell-free DNA molecule in the particular viral genome, the sizes of the set of cell-free DNA molecules forming a size distribution, the sample being the biological sample or a different sample including a mixture of cell-free DNA molecules from the genome of the subject and from the one or more other genomes;

determining a statistical value of the size distribution;

comparing the statistical value to a reference size value determined from the at least two cohorts of other subjects;

determining a second classification of whether the subject has the first condition based on the comparing of the statistical value to the reference size value; and

determining a final classification using the first classification and the second classification.

24. The method of claim 1, further comprising:

determining an amount of cell-free DNA molecules derived from the particular viral genome in a sample, the sample being the biological sample or a different sample including a mixture of cell-free DNA molecules from the genome of the subject and from the one or more other genomes;

comparing the amount to a reference value determined from the at least two cohorts of other subjects;

determining a second classification of whether the subject has the first condition based on the comparing of the amount to the reference value; and

determining a final classification using the first classification and the second classification.

25. The method of claim 1, further comprising:

responsive to the first classification being that the subject has the first condition, providing a treatment to the subject to improve the first condition.

26. A computer product comprising a computer readable medium storing a plurality of instructions for controlling a computer system to perform operations of the methods of any one of claims 1-25.

27. A system comprising:

the computer product of claim 26; and

one or more processors for executing instructions stored on the computer readable medium.

28. A system comprising means for performing the methods of any one of claims 1-25.

29. A system configured to perform the methods of any one of claims 1-25.

30. A system comprising modules that respectively perform the steps of the methods of any one of claims 1-25.