

condition, or the progression of a condition over time. For example, a method of the present disclosure may be used to determine a stage of cancer in a subject, or the progression of cancer in a subject over time (e.g., using samples obtained from a subject at different times).

5 **[0003a]** In one aspect, the present disclosure refers to a method of analyzing a biological sample, including a mixture of cell-free nucleic acid molecules, to determine a level of a cancer associated with a virus in a subject from which the biological sample is obtained, the mixture including nucleic acid molecules from the subject and potentially nucleic acid molecules from the virus, wherein the biological sample is blood, plasma, serum, or urine, the method comprising: for each of a plurality of nucleic acid molecules in the biological sample:
10 measuring a size of the nucleic acid molecule; determining that the nucleic acid molecule is from a reference genome, the reference genome corresponding to the virus; determining a statistical value of a size distribution of the plurality of nucleic acid molecules from the reference genome; and determining the level of the cancer in the subject by processing the statistical value against a cutoff value.

15 **[0003b]** In another aspect, the present disclosure refers to a method of screening a biological sample, including a mixture of cell-free nucleic acid molecules, for a cancer associated with a virus, the biological sample obtained from a subject that is asymptomatic for the cancer, the mixture including nucleic acid molecules from the subject and potentially nucleic acid molecules from the virus, wherein the biological sample is blood, plasma, serum,
20 or urine, the method comprising: receiving a plurality of sequence reads obtained from a sequencing of the mixture of cell-free nucleic acid molecules; determining an amount of the plurality of sequence reads aligning to a reference genome, the reference genome corresponding to the virus; and comparing the amount of sequence reads aligning to the reference genome to a cutoff value, thereby screening for the cancer, the cutoff value
25 differentiating subjects that transiently or persistently have the virus but not the cancer and subjects that have the cancer.

[0003c] In another aspect, the present disclosure refers to a method of analyzing a biological sample, including a mixture of cell free nucleic acid molecules, to determine a level of cancer associated with a virus in a subject from which the biological sample is
30 obtained, the mixture including nucleic acid molecules from the subject and potentially nucleic acid molecules from the virus, wherein the biological sample is blood, plasma, serum, or urine, the method comprising: performing a first assay, wherein the first assay comprises analyzing a first plurality of cell-free nucleic acid molecules from a first biological sample of the subject to determine a first amount of the first plurality of cell-free nucleic acid molecules
35 aligning to a reference genome, the reference genome corresponding to the virus; and performing a second assay, wherein the second assay comprises measuring a size of each of a second plurality of cell-free nucleic acid molecules in a second biological sample; and

determining a second amount of a size distribution of the second plurality of cell-free nucleic acid molecules from the reference genome, the second amount being a statistical value of the size distribution of the second plurality of cell-free nucleic acid molecules from the reference genome; comparing the first amount to a first cutoff; comparing the second amount to a second cutoff; and
5 determining the level of the cancer in the subject based on the comparing of the first amount to the first cutoff and the comparing of the second amount to the second cutoff.

[0003d] In yet another aspect, the present disclosure refers to a method of analyzing a biological sample, including a mixture of cell free nucleic acid molecules, to determine a level of cancer associated with a virus in a subject from which the biological sample is
10 obtained, the mixture including nucleic acid molecules from the subject and potentially nucleic acid molecules from the virus, wherein the biological sample is blood, plasma, serum, or urine, the method comprising: analyzing a first plurality of cell-free nucleic acid molecules from the biological sample of the subject, wherein the analyzing a cell-free nucleic acid comprises determining a genomic position in a reference genome corresponding to at least
15 one end of the cell-free nucleic acid molecule, the reference genome corresponding to the virus; determining a first amount of the first plurality of cell-free nucleic acid molecules that end within one or more first windows, each first window comprising at least one of a first set of genomic positions at which ends of cell-free nucleic acid molecules are present at a rate above a first threshold in subjects with the cancer associated with the virus; computing a
20 relative abundance of the first plurality of cell-free nucleic acid molecules ending within one of the one or more first windows by normalizing the first amount using a second amount of the first plurality of cell-free nucleic acid molecules from the biological sample, wherein the second amount of the first plurality of cell-free nucleic acid molecules includes cell-free nucleic acid molecules ending at a second set of genomic positions outside of the one or more
25 first windows including the first set of genomic positions; and determining the level of cancer in the subject by processing the relative abundance against a cutoff value.

[0003e] In yet 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 the method as disclosed herein.

30 **[0003f]** In yet another aspect, the present disclosure refers to a computer system including one or more processors that are configured to perform the method as disclosed herein.

WHAT IS CLAIMED IS:

1. A method of analyzing a biological sample, including a mixture of cell-free nucleic acid molecules, to determine a level of a cancer associated with a virus in a subject from which the biological sample is obtained, the mixture including nucleic acid molecules from the subject and potentially nucleic acid molecules from the virus, wherein the biological sample is blood, plasma, serum, or urine, the method comprising:

for each of a plurality of nucleic acid molecules in the biological sample:

measuring a size of the nucleic acid molecule;

determining that the nucleic acid molecule is from a reference genome, the reference genome corresponding to the virus;

determining a statistical value of a size distribution of the plurality of nucleic acid molecules from the reference genome; and

determining the level of the cancer in the subject by processing the statistical value against a cutoff value.

2. The method of claim 1, wherein the statistical value is an average, mode, median, or mean of the size distribution.

3. The method of claim 1, wherein the statistical value is a percentage of the plurality of nucleic acid molecules in the biological sample from the reference genome that are below a size threshold.

4. The method of claim 3, wherein the size threshold is 150 bp.

5. The method of claim 3, wherein the subject is determined to be positive for the cancer when the statistical value is below the cutoff value.

6. The method of claim 1, wherein the statistical value includes a ratio of:
a first amount the plurality of nucleic acid molecules in the biological sample from the reference genome that are within a first size range; and

a second amount the plurality of nucleic acid molecules in the biological sample from the reference genome that are within a second size range that is different than the first size range.

7. The method of claim 1, wherein the statistical value includes a ratio of:

a first proportion of sequence reads of nucleic acid molecules that align to the reference genome of the virus with the size within a given range; and

a second proportion of sequence reads of nucleic acid molecules that align to a human reference genome with the size within the given range.

8. The method of claim 7, wherein the cutoff value is about 7, about 8, about 9, or about 10.

9. The method of claim 7, wherein the given range is about 80 to about 110 base pairs, about 50 to about 75 base pairs, about 60 to about 90 base pairs, about 90 to about 120 base pairs, about 120 to about 150 base pairs, or about 150 to about 180 base pairs.

10. The method of claim 7, wherein the statistical value is an inverse of the ratio.

11. The method of claim 10, wherein the cutoff value is about 0.11, 0.12, 0.13, or 0.14.

12. The method of claim 1, wherein measuring the size of the nucleic acid molecule and determining that the nucleic acid molecule is from the reference genome includes:

receiving one or more sequence reads that include both ends of the nucleic acid molecule, thereby obtaining a plurality of sequence reads from a sequencing of the mixture of cell free nucleic acid molecules;

aligning the one or more sequence reads to the reference genome to obtain one or more aligned locations; and

using the one or more aligned locations to determine the size of the nucleic acid molecule.

13. The method of claim 1, wherein a positive predictive value (PPV) of the determining the level of the cancer is at least 6%, 7%, or 8%, wherein a sensitivity of the determining the level of the cancer is at least 95%, 96%, or 97%, and/or wherein a specificity of the determining the level of the cancer is at least 95%, 96%, 97%, or 98%.

14. The method of claim 1, wherein the size is a length, a molecular mass, or a measured parameter that is proportional to the length.

15. A method of screening a biological sample, including a mixture of cell-free nucleic acid molecules, for a cancer associated with a virus, the biological sample obtained from a subject that is asymptomatic for the cancer, the mixture including nucleic acid molecules from the subject and potentially nucleic acid molecules from the virus, wherein the biological sample is blood, plasma, serum, or urine, the method comprising:

receiving a plurality of sequence reads obtained from a sequencing of the mixture of cell-free nucleic acid molecules;

determining an amount of the plurality of sequence reads aligning to a reference genome, the reference genome corresponding to the virus; and

comparing the amount of sequence reads aligning to the reference genome to a cutoff value, thereby screening for the cancer, the cutoff value differentiating subjects that transiently or persistently have the virus but not the cancer and subjects that have the cancer.

16. The method of claim 15, further comprising:

determining a level of the cancer in the subject based on the comparing of the amount of sequence reads to the cutoff value.

17. The method of claim 15, wherein the amount of sequence reads aligning to the reference genome includes a proportion of sequence reads aligned to the reference genome relative to a total number of sequence reads.

18. The method of claim 17, wherein the total number of sequence reads is a sum of the sequence reads that aligned to the reference genome corresponding to the virus and the sequence reads that aligned to a human genome.

19. The method of claim 15, wherein the amount of sequence reads aligning to the reference genome includes a number of sequence reads aligned to the reference genome per volume of the biological sample analyzed.

20. The method of claim 15, wherein the cutoff value is determined from training samples having a known classification for the cancer.

21. The method of claim 20, wherein the cutoff value is selected using (1) a value below a lowest amount of sequence reads aligning to the reference genome for the training samples classified as having the cancer; (2) a specified number of standard

deviations from a mean amount of sequence reads aligning to the reference genome for the training samples classified as having the cancer; or (3) a specificity and a sensitivity for determining a correct classification of the training samples.

22. The method of one of claim 12 or claim 15, further comprising:
performing the sequencing of the mixture of cell-free nucleic acid molecules
to obtain the plurality of sequence reads.

23. The method of claim 22, further comprising:
performing real-time polymerase chain reaction (PCR) of the biological
sample or a different biological sample obtained from the subject contemporaneously as the
biological sample, thereby determining a quantity of nucleic acid molecules from the virus;
comparing the quantity to a quantity threshold; and
when the quantity is above the quantity threshold, performing the sequencing
of the mixture of cell-free nucleic acid molecules.

24. The method of claim 22, wherein the sequencing includes random
sequencing.

25. The method of claim 15, further comprising:
comparing the amount of sequence reads aligning to the reference genome to
another cutoff value, the other cutoff value differentiating between subjects that transiently
have the virus but not the cancer and subjects that persistently have the virus but not the
cancer.

26. A method of analyzing a biological sample, including a mixture of
cell-free nucleic acid molecules, to determine a level of a cancer associated with a virus in a
subject from which the biological sample is obtained, the mixture including nucleic acid
molecules from the subject and potentially nucleic acid molecules from the virus, wherein the
biological sample is blood, plasma, serum, or urine, the method comprising:

performing a first assay, wherein the first assay comprises analyzing a first
plurality of cell-free nucleic acid molecules from a first biological sample of the subject to
determine a first amount of the first plurality of cell-free nucleic acid molecules aligning to a
reference genome, the reference genome corresponding to the virus; and

performing a second assay, wherein the second assay comprises

measuring a size of each of a second plurality of cell-free nucleic acid

molecules in a second biological sample; and
determining a second amount of a size distribution of the second plurality of cell-free nucleic acid molecules from the reference genome, the second amount being a statistical value of the size distribution of the second plurality of cell-free nucleic acid molecules from the reference genome;
comparing the first amount to a first cutoff;
comparing the second amount to a second cutoff; and
determining the level of the cancer in the subject based on the comparing of the first amount to the first cutoff and the comparing of the second amount to the second cutoff.

27. The method of claim 26, wherein the first assay includes real-time polymerase chain reaction (PCR).

28. The method of claim 1, 15, or 26, wherein the cancer is selected from a group consisting of nasopharyngeal cancer, head and neck squamous cell carcinoma, cervical cancer, and hepatocellular carcinoma.

29. The method of claim 26, wherein the first biological sample and the second biological sample are from a same blood sample.

30. The method of claim 26, wherein the first biological sample and the second biological sample are the same sample, and wherein the second plurality of cell-free nucleic acid molecules is the first plurality of cell-free nucleic acid molecules.

31. The method of claim 26, wherein the second assay is performed in response to the first amount being above the first cutoff.

32. The method of claim 26, wherein the subject is determined to have the cancer only if the first amount exceeds the first cutoff and the second amount exceeds the second cutoff.

33. The method of claim 26, wherein the second amount is of nucleic acid molecules having the size within a given range and aligning to the reference genome.

34. The method of claim 33, further comprising normalizing the second amount using a third amount of cell-free nucleic acid molecules having the size within a different range and aligning to the reference genome.

35. The method of claim 33, further comprising normalizing the second amount using a third amount of cell-free nucleic acid molecules having the size within the given range and aligning to an autosomal genome.

36. The method of claim 26, wherein a positive predictive value of determining the level of the cancer is at least 15%, at least 17%, or at least 19%, wherein a sensitivity of determining the level of the cancer is at least 95%, at least 96%, or at least 97%, and/or wherein a specificity of determining the level of the cancer is at least 95%, at least 96%, at least 97%, at least 98%, or at least 99%.

37. The method of any one of claims 1, 15, or 26, further comprising enriching the biological sample for nucleic acid molecules from the virus.

38. The method of claim 37, wherein the enriching the biological sample for nucleic acid molecules from the virus includes using capture probes that bind to a portion of, or an entire genome of, the virus.

39. The method of claim 37, further comprising:
enriching the biological sample for nucleic acid molecules from a portion of a human genome.

40. A method of analyzing a biological sample, including a mixture of cell-free nucleic acid molecules, to determine a level of cancer associated with a virus in a subject from which the biological sample is obtained, the mixture including nucleic acid molecules from the subject and potentially nucleic acid molecules from the virus, wherein the biological sample is blood, plasma, serum, or urine, the method comprising:

analyzing a first plurality of cell-free nucleic acid molecules from the biological sample of the subject, wherein the analyzing a cell-free nucleic acid comprises determining a genomic position in a reference genome corresponding to at least one end of the cell-free nucleic acid molecule, the reference genome corresponding to the virus;

determining a first amount of the first plurality of cell-free nucleic acid

molecules that end within one or more first windows, each first window comprising at least one of a first set of genomic positions at which ends of cell-free nucleic acid molecules are present at a rate above a first threshold in subjects with the cancer associated with the virus;

computing a relative abundance of the first plurality of cell-free nucleic acid molecules ending within one of the one or more first windows by normalizing the first amount using a second amount of the first plurality of cell-free nucleic acid molecules from the biological sample, wherein the second amount of the first plurality of cell-free nucleic acid molecules includes cell-free nucleic acid molecules ending at a second set of genomic positions outside of the one or more first windows including the first set of genomic positions; and

determining the level of cancer in the subject by processing the relative abundance against a cutoff value.

41. The method of claim 40, wherein processing the relative abundance against the cutoff value includes:

determining whether the relative abundance is greater than the cutoff value.

42. The method of claim 40, further comprising:

determining the second amount of the first plurality of cell-free nucleic acid molecules that end within one of second windows, each second window comprising at least one of the second set of genomic positions at which ends of cell-free nucleic acid molecules are present at a rate above a second threshold in subjects without the cancer resulting from the virus,

wherein normalizing the first amount includes computing the relative abundance using the first amount and the second amount.

43. The method of claim 42, further comprising identifying the second set of genomic positions, wherein the identifying comprises:

analyzing the cell-free nucleic acid molecules of a reference sample from a reference subject that does not have the cancer, wherein analyzing each of the cell-free nucleic acid molecules of the reference sample comprises:

determining the genomic position in the reference genome corresponding to at least one end of the cell-free nucleic acid molecule.

44. The method of claim 43, wherein the reference subject is healthy.

45. The method of claim 42, wherein the relative abundance comprises a ratio of the first amount and the second amount.

46. The method of claim 40, further comprising identifying the first set of genomic positions at which ends of cell-free nucleic acid molecules occur at the rate above the first threshold.

47. The method of claim 46, wherein identifying the first set of genomic positions comprises:

analyzing a second plurality of cell-free nucleic acid molecules from at least one first additional sample to identify ending positions of the second plurality of cell-free nucleic acid molecules, wherein the at least one first additional sample is known to have the cancer associated with the virus and is of a same sample type as the biological sample;

for each genomic window of a plurality of genomic windows:

computing a corresponding number of the second plurality of cell-free nucleic acid molecules ending on the genomic window; and

comparing the corresponding number to a reference value to determine whether the rate of cell-free nucleic acid molecules ending on one or more genomic positions within the genomic window is above the first threshold.

48. The method of claim 47, wherein a first genomic window of the plurality of genomic windows has a width of at least one genomic position, and wherein each of the genomic positions within the first genomic window are identified as having the rate of cell-free nucleic acid molecules ending on the genomic position be above the first threshold when the corresponding number exceeds the reference value.

49. The method of claim 47, wherein each genomic position of the first set of genomic positions has at least a specified number of cell-free nucleic acid molecules of the second plurality of cell-free nucleic acid molecules ending on the genomic position.

50. The method of claim 47, wherein the reference value is an expected number of cell-free nucleic acid molecules ending within the genomic window according to a probability distribution and an average length of cell-free nucleic acid molecules in the at least one first additional sample.

51. The method of claim 47, wherein the reference value comprises a measured number of cell-free nucleic acid molecules ending within the genomic window, the measured number determined from a third plurality of cell-free nucleic acid molecules of at least one second additional sample identified as not having the cancer.

52. The method of any one of claims 1, 15, 26, and 40, wherein the level of the cancer is selected from a group consisting of: an amount of tumor tissue in the subject, a size of a tumor in the subject, a stage of the tumor in the subject, a tumor load in the subject, and presence of tumor metastasis in the subject.

53. The method of any one of claims 1, 15, 26, and 40, wherein the virus comprises EBV DNA, HPV DNA, HBV DNA, HCV nucleic acids, or fragments thereof.

54. The method of any one of claims 1, 15, 26, and 40, further comprising that (a) a therapeutic intervention is to be provided to the subject based on the level of the cancer, or (b) imaging is to be performed on the subject based on the level of the cancer.

55. The method of any one of claims 1, 15, 26, and 40, wherein the cell-free nucleic acid molecules are deoxyribonucleic acid (DNA) molecules.

56. A computer product comprising a computer readable medium storing a plurality of instructions for controlling a computer system to perform the method of any one of claims 1-55.

57. A computer system including one or more processors that are configured to perform the method of any one of claims 1-55.