

CLAIMS

1. A computer product comprising a non-transitory computer readable medium storing a plurality of instructions for controlling a computer system to perform operations of a method of screening a pathogen-associated disorder in a subject, the method comprising:

receiving data from a first assay performed at a first time point that comprises determining a characteristic of cell-free nucleic acid molecules from a pathogen in a biological sample of the subject, wherein the characteristic of the cell-free nucleic acid molecules from the pathogen comprises amount, methylation status, variant pattern, fragment size, or relative abundance as compared to cell-free nucleic acid molecules from the subject in the biological sample, and wherein the characteristic indicates a risk for the subject to develop the pathogen-associated disorder; and

determining, based on the characteristic, a second time point at which a second assay is performed to screen for the pathogen-associated disorder in the subject, wherein an interval between the first time point and the second time point inversely correlates with the risk.

2. A computer product comprising a non-transitory computer readable medium storing a plurality of instructions for controlling a computer system to perform operations of a method of prognosticating a pathogen-associated disorder in a subject, the method comprising:

receiving data from a first assay that comprises determining a characteristic of cell-free nucleic acid molecules from a pathogen in a biological sample of the subject, wherein the characteristic of the cell-free nucleic acid molecules from the pathogen comprises amount, methylation status, variant pattern, fragment size, or relative abundance as compared to cell-free nucleic acid molecules from the subject in the biological sample; and

generating a report indicative of a risk for the subject to develop the pathogen-associated disorder based on the characteristic of the cell-free nucleic acid molecules from the pathogen, and one or more factors of age of the subject, smoking habit of the subject, family history of the pathogen-associated disorder of the subject, genotypic factors of the subject, ethnicity of the subject, or dietary history of the subject.

3. The computer product of claim 1, wherein result of the first assay does not result in a medical treatment of the subject for the pathogen-associated disorder, optionally wherein the

medical treatment comprises treatment with therapeutic agents, radiotherapy, or surgical treatment.

4. The computer product of claim 1 or 3, wherein the subject is diagnosed as not having the pathogen-associated disorder before the determining a second time point by a clinical diagnostic examination that has a false positive rate below 1%, optionally wherein the clinical diagnostic examination comprises physical examination, invasive biopsy, endoscopy, magnetic resonance imaging, positive emission tomography, computed tomography, or x-ray imaging, wherein said invasive biopsy optionally comprises histological analysis, cytological analysis, or cellular nucleic acid analysis.

5. The computer product of any one of claims 1, or 3 to 4, wherein the interval is at least about 2 months, 4 months, 6 months, 8 months, 10 months, or 12 months.

6. The computer product of any one of claims 1 to 5, further comprising performing the first assay, said first assay comprising:

A) (i) obtaining a first biological sample from the subject; and

(ii) measuring a first amount of cell-free nucleic acid molecules from the pathogen in the first biological sample, such as measuring a first percentage of the cell-free nucleic acid molecules from the pathogen in the first biological sample, optionally wherein the measuring the first amount comprises:

- measuring a copy number of the cell-free nucleic acid molecules from the pathogen in the first biological sample, and/or
- polymerase chain reaction (PCR) or quantitative PCR (qPCR);

B) determining a methylation status of the cell-free nucleic acid molecules from the pathogen in the biological sample, optionally wherein the determining the methylation status comprises treatment of the cell-free nucleic acid molecules in the biological sample with a methylation-sensitive restriction enzyme or bisulfite, or comprises performing a methylation-aware sequencing of cell-free nucleic acids in the biological sample of the subject, such as one that comprises bisulfite conversion of unmethylated cytosine to uracil or that comprises treatment with a methylation-sensitive restriction enzyme;

C) determining a fragment size distribution of the cell-free nucleic acid molecules from the pathogen in the biological, optionally comprising performing sequencing on cell-free nucleic acid molecules in the biological sample, and determining a fragment size of the cell-free nucleic acid molecules from the pathogen in the biological sample based on sequence reads mapped to the reference genome of the pathogen; or

D) determining a variant pattern of the cell-free nucleic acid molecules from the pathogen in the biological sample, optionally wherein:

- the determining the variant pattern comprises performing sequencing on cell-free nucleic acid molecules in the biological sample, and determining the variant pattern of the cell-free nucleic acid molecules from the pathogen in the biological sample based on sequence reads mapped to the reference genome of the pathogen, and/or
- the variant pattern of the cell-free nucleic acid molecules from the pathogen comprises single nucleotide variations.

7. The computer product of claim 6, wherein the first assay comprises A) and wherein the first assay further comprises: (iii) if the first amount is above a threshold, obtaining a second biological sample from the subject, for example about 4 weeks after the first biological sample, and measuring a second amount of cell-free nucleic acid molecules from the pathogen in the second biological sample, optionally wherein:

- the interval between the first time point and the second time point is shorter if both the first amount and the second copy number are above the threshold as compared to an interval if the second amount is below the threshold,
- the interval between the first time point and the second time point is longer if the first amount is below the threshold as compared to an interval if the first amount is above the threshold,
- the interval between the first time point and the second time point is about 1 year if both the first amount and the second amount are above the threshold,
- the interval between the first time point and the second time point is about 2 years if the second amount is below the threshold, and/or
- the interval between the first time point and the second time point is about 4 years if the first amount is below the threshold.

8. The computer product of claim 6 or 7, wherein the first assay comprises determining the methylation status, the fragment size distribution, or the variant pattern of the cell-free nucleic acid molecules from the pathogen in the biological sample.

9. The computer product of claim 1, further comprising performing the second assay at the second time point, optionally wherein:

- i) the second assay is same as the first assay, or
- ii) the second assay comprises an assay of cell-free nucleic acid molecules from the subject, an invasive biopsy of the subject, endoscopic examination of the subject, or magnetic resonance imaging examination of the subject.

10. A computer product comprising a non-transitory computer readable medium storing a plurality of instructions for controlling a computer system to perform operations of a method of analyzing nucleic acid molecules from a biological sample of a subject, the method comprising:

obtaining, in a computer system, sequence reads of cell-free nucleic acid molecules from the biological sample of the subject, such as biological sample selected from the group consisting of: whole blood, blood plasma, blood serum, urine, cerebrospinal fluid, buffy coat, vaginal fluid, vaginal flushing fluid, saliva, oral rinse fluid, nasal flushing fluid, a nasal brush sample and a combination thereof, wherein the biological sample comprises cell-free nucleic acid molecules from the subject and potentially from a pathogen;

aligning, in the computer system, the sequence reads of the cell-free nucleic acid molecules to a reference genome of the pathogen; and

identifying, in the computer system, a variant pattern of the cell-free nucleic acid molecules from the pathogen, for example comprising single nucleotide variations, wherein the variant pattern characterizes a nucleotide variant of the sequence reads mapped to the reference genome of the pathogen at each of a plurality of variant sites on the reference genome of the pathogen, wherein the plurality of variant sites comprises at least 30 sites across the reference genome of the pathogen, and wherein the variant pattern indicates a status of, or a risk for, a pathogen-associated disorder in the subject.

11. The computer product of claim 10, wherein:

- the plurality of variant sites comprises at least 40, at least 50, at least 60, at least 70, at least 80, at least 90, at least 100, at least 200, at least 300, at least 400, at least 500, at least 600, at least 700, at least 800, at least 900, at least 1000, at least 1100, or at least 1200 sites across the reference genome of the pathogen;
- the plurality of variant sites consists of all sites at which the sequence reads mapped to the reference genome of the pathogen have a different nucleotide variant than the reference genome of the pathogen;
- the aligning the sequence reads is configured to allow a maximum mismatch of 10, 9, 8, 7, 6, 5, 4, 3, 2, or 1 bases between the sequence reads mapped to the reference genome of the pathogen and the reference genome of the pathogen; or is configured to allow a maximum mismatch of 2 bases between the sequence reads mapped to the reference genome of the pathogen and the reference genome of the pathogen.

12. The computer product of claim 10 or 11, further comprising:

diagnosing, prognosticating, or monitoring the pathogen-associated disorder in the subject based on the variant pattern of the sequence reads mapped to the reference genome of the pathogen.

13. The computer product of any one of claims 10 to 12, wherein the pathogen is a virus.

14. The computer product of claim 13, wherein the virus is Epstein-Barr virus (EBV), optionally wherein the pathogen-associated disorder comprises nasopharyngeal cancer, NK cell lymphoma, Burkitt's lymphoma, post-transplant lymphoproliferative disorders, or Hodgkin's lymphoma, and optionally wherein:

- i) the variant pattern of the cell-free nucleic acid molecules from the pathogen characterizes nucleotide variant of the sequence reads mapped to the referenced genome of the pathogen at each of the plurality of variant sites that comprises at least 30, 40, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, or 600 sites selected from genomic sites as set forth in **Table 6** relative to EBV reference genome (AJ507799.2), optionally wherein the plurality of variant sites comprises a genomic site as set forth in **Table 6** relative to EBV reference genome (AJ507799.2),
- ii) the variant pattern of the cell-free nucleic acid molecules from the pathogen characterizes nucleotide variant of the sequence reads mapped to the referenced genome

of the pathogen at each of the plurality of variant sites that are randomly selected from genomic sites as set forth in **Table 6** relative to EBV reference genome (AJ507799.2), or
iii) the variant pattern of the cell-free nucleic acid molecules from the pathogen characterizes nucleotide variant of the sequence reads mapped to the referenced genome of the pathogen at each of the plurality of variant sites that comprise at least 30, 40, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, or 600 sites randomly selected from genomic sites as set forth in **Table 6** relative to EBV reference genome (AJ507799.2).

15. The computer product of claim 13, wherein the virus is:

- human papillomavirus (HPV), optionally the pathogen-associated disorder comprising cervical cancer, oropharyngeal cancer, or head and neck cancers; or
- hepatitis B virus (HBV), optionally the pathogen-associated disorder comprising cirrhosis or hepatocellular carcinoma (HCC).

16. The computer product of any one of claims 10 to 15, wherein the variant pattern indicates a status of a pathogen-associated disorder in the subject, wherein the status of the pathogen-associated disorder comprises a presence of the pathogen-associated disorder in the subject, an amount of tumor tissue in the subject, a size of the tumor tissue in the subject, a stage of tumor in the subject, a tumor load in the subject, or a presence of tumor metastasis in the subject.

17. The computer product of claim 6 wherein the first assay comprises D), or the method of any one of claims 6 to 11, wherein the determining or identifying the variant pattern comprises:

determining a similarity level between the sequence reads mapped to the reference genome of the pathogen and a disorder-related reference genome of the pathogen, optionally wherein the disorder-related reference genome of the pathogen comprises a genome of the pathogen identified in a diseased tissue.

18. The computer product of claim 17, wherein the determining the similarity level comprises:

segregating the reference genome of the pathogen into a plurality of bins, optionally the plurality of bins having a length of about 100, 200, 300, 400, 500, 600, 700, 800, 900, or 1000 bp; and

determining a similarity index for each of the plurality of bins against the disorder-related reference genome of the pathogen, wherein the similarity index correlates with a proportion of the variant sites, within the respective bin, at which at least one of the sequence reads mapped to the reference genome of the pathogen has a same nucleotide variant as the disorder-related reference genome of the pathogen, optionally wherein the disorder-related reference genome of the pathogen comprises a plurality of disorder-related reference genomes of the pathogen, and wherein the determining the similarity level comprises:

determining a respective similarity index for each of the plurality of bins against each of the plurality of disorder-related reference genomes of the pathogen; and

determining a bin score for each of the plurality of bins based on a proportion of the plurality of disorder-related reference genomes, against which the respective similarity index within the respective bin is above a cutoff value, optionally said cutoff value being about 0.9.

19. The computer product of any one of claims 1 to 18, further comprising: calculating a risk score for the subject to develop the pathogen-associated disorder using a classifier applied to a data input comprising the variant pattern of the cell-free nucleic acid molecules from the pathogen, wherein the classifier is configured to apply a function to the data input comprising the variant pattern of the cell-free nucleic acid molecules from the pathogen to generate an output comprising the risk score that evaluates the risk for the subject to develop the disorder, optionally wherein:

- (a) - the classifier is trained with a labeled dataset, and/or
- (b) - the classifier comprises a mathematical model using Naïve Bayes model, logistics regression, random forest, decision tree, gradient boosting tree, neural network, deep learning, linear/kernel support vector machine (SVM), linear/non-linear regression, or linear discriminative analysis.

20. A system comprising:
the computer product of any one of claims 1-19; and
one or more processors for executing instructions stored on the computer readable medium.