

CANCER DETECTION METHOD, KIT, AND SYSTEM

CROSS-REFERENCE TO RELATED APPLICATION

[0001] The present application claims the benefit of U.S. Provisional Application No. 63/208,506 filed on June 9, 2021, whose disclosure is hereby incorporated by reference in its entirety.

REFERENCE TO SEQUENCE LISTING SUBMITTED ELECTRONICALLY

[0002] The content of the electronically submitted sequence listing, file name Top_miRNA_Seq.txt, size 15,063 bytes, and date of creation May 31, 2022, filed herewith, is incorporated herein by reference in its entirety.

FIELD OF THE INVENTION

[0003] The present invention relates generally to the technical field of disease screening, detection and diagnosis, and more specifically relates to a method, a kit, a system, and a non-transitory storage medium for the detection of one or multiple human cancers.

BACKGROUND

[0004] Despite the recent rapid development of diagnosis and treatment technologies, cancers have remained largely as a challenging and potentially lethal disease for humans. It is well appreciated that detection of cancer at early stages is critical to decrease cancer-related mortality as treatment is more likely to be successful at early stages. There is an urgent unmet need to develop a test capable of detecting multiple cancer types early and simultaneously, ideally noninvasive, such as a blood test, which has become the cornerstone of the so-called multi-cancer early detection (MCED) paradigm. Such a MCED test often requires very high specificity, preferably >99%, to ensure minimum false positives in order to be able to screen the at-risk general population.

[0005] Molecules such as microRNAs (i.e. miRNAs) may serve as biomarkers for MCED. miRNAs are small single-stranded non-coding RNA molecules of an average of 22 nucleotides long encoded by their corresponding genes in the human genome. The miRNAs function in negative post-transcriptional regulation of gene expression primarily by binding with complementary sequences in the 3' untranslated region (3' UTR) of mRNA molecules. miRNAs appear to regulate more than 50% human genes, and abnormal expression of miRNAs has been implicated in many human cancers. miRNAs are also abundant as extracellular circulating molecules released into circulation by tumor cells either through cell death or by exosome-

mediated signaling. Combined with its remarkable stability in the blood and other body fluids, circulating cell-free miRNAs have the potential to serve as noninvasive biomarkers for cancer screening and diagnosis.

5

SUMMARY

[0006] The present disclosure provides a multi-cancer detection approach (i.e. method, kit, and system) by means of an miRNA biomarker set consisting of at least one miRNA biomarker. The approach is substantially based on the expression profile of the miRNA biomarker set, which can be determined from a biological sample obtained from a human subject. Such a biological sample can notably be a liquid biopsy sample including a blood sample, a serum sample, a plasma sample, a urine sample, a saliva sample, or a spatum sample to thereby allow a non-invasive or minimum-invasive detection for the cancer. The approach can be employed to accurately and reliably detect whether a human subject has one of the cancers including lung cancer, biliary tract cancer, bladder cancer, colorectal cancer, esophageal cancer, gastric cancer, glioma cancer, liver cancer, pancreatic cancer, prostate cancer, ovarian cancer, and sarcoma.

[0007] In a first aspect, a method for detecting a cancer from a biological sample obtained from a subject is provided. The method substantially includes the following three steps (1)-(3):

[0008] Step (1): determining an expression profile of an miRNA biomarker set consisting of at least one miRNA from the biological sample. Herein, the miRNA biomarker set comprises hsa-miR-5100.

[0009] Step (2): calculating a diagnostic index of the biological sample based on the expression profile of the miRNA biomarker set. The diagnostic index is calculated based on:

$$\text{diagnostic index} = \sum_{i=1}^n t_i * \text{miRNA}_i; \quad (\text{I})$$

where n is the total number of miRNAs in the miRNA biomarker set, miRNA_i is the expression level of the i^{th} miRNA in the miRNA biomarker set, where i is an integer greater than zero and smaller than or equal to n ; and t_i is a weight for the i^{th} miRNA.

[0010] Step (3): classifying the subject as having the cancer or not based on the value of the calculated diagnostic index. If the calculated diagnostic index is greater than or equal to a pre-determined threshold, the subject is classified as having the cancer; or if otherwise the subject is classified as not having the cancer.

[0011] It is further configured such that the method is capable of achieving diagnostic accuracy having an AUC value greater than approximately 0.780.

[0012] As used herein, the expression profile of an miRNA biomarker set is substantially a dataset containing expression level data that has been determined for each and every miRNA

member contained in the miRNA biomarker set.

[0013] The term “pre-determined threshold” is referred to as a cut-point value of the diagnostic index that can be used to determine with a given specificity/sensitivity if a subject has the cancer type or not. It is typically pre-determined based on an existing dataset comprising a range of diagnostic index values that have been obtained and calculated for an existing population of subjects known to have, and/or known to be absent of, the disease. For example, in the EXAMPLE 1 provided below, when the miRNA biomarker set consists of any one of the top 100 miRNAs (corresponding to SEQ ID NOS: 1-100), the AUC can reach a level that is greater than 0.780 (i.e. for hsa-miR-1238-5p), and can even reach approximately 0.999 (i.e. for the top 4 miRNAs of hsa-miR-5100, hsa-miR-1343-3p, hsa-miR-1290 and hsa-miR-4787-3p) (See Table 1).

[0014] According to some embodiments of the method, the miRNA biomarker set further comprises, in addition to hsa-miR-5100 (corresponding to SEQ ID NO: 1), one or more of the other 99 miRNAs listed in Table 1, i.e. hsa-miR-1343-3p, hsa-miR-1290, hsa-miR-4787-3p, hsa-miR-6877-5p, hsa-miR-17-3p, hsa-miR-6765-5p, hsa-miR-1268b, hsa-miR-4258, hsa-miR-451a, hsa-miR-1228-5p, hsa-miR-8073, hsa-miR-4454, hsa-miR-187-5p, hsa-miR-4286, hsa-miR-6746-5p, hsa-miR-663b, hsa-miR-6075, hsa-miR-5001-5p, hsa-miR-6789-5p, hsa-miR-4513, hsa-miR-3192-5p, hsa-miR-8060, hsa-miR-668-5p, hsa-miR-1268a, hsa-miR-1273g-3p, hsa-miR-4706, hsa-miR-124-3p, hsa-miR-1260b, hsa-miR-4740-5p, hsa-miR-320b, hsa-miR-7977, hsa-miR-29b-3p, hsa-miR-4708-3p, hsa-miR-4525, hsa-miR-92b-3p, hsa-miR-4257, hsa-miR-4727-3p, hsa-miR-92a-3p, hsa-miR-663a, hsa-miR-6787-5p, hsa-miR-3131, hsa-miR-6802-5p, hsa-miR-654-5p, hsa-miR-6511b-5p, hsa-miR-29b-1-5p, hsa-miR-4417, hsa-miR-4736, hsa-miR-6840-3p, hsa-miR-4710, hsa-miR-4635, hsa-miR-296-3p, hsa-miR-1199-5p, hsa-miR-7975, hsa-miR-4480, hsa-miR-3648, hsa-miR-371a-5p, hsa-miR-4771, hsa-miR-6717-5p, hsa-miR-1254, hsa-miR-1246, hsa-miR-23b-3p, hsa-miR-320a, hsa-miR-4687-5p, hsa-miR-191-5p, hsa-miR-320c, hsa-miR-6131, hsa-miR-4515, hsa-miR-342-5p, hsa-miR-4718, hsa-miR-23a-3p, hsa-miR-4455, hsa-miR-211-3p, hsa-miR-3122, hsa-miR-103a-3p, hsa-miR-4429, hsa-miR-920, hsa-miR-3194-3p, hsa-miR-4754, hsa-miR-1238-5p, hsa-miR-3191-3p, hsa-miR-4755-3p, hsa-miR-3688-5p, hsa-miR-4529-5p, hsa-miR-6861-5p, hsa-miR-1469, hsa-miR-619-5p, hsa-miR-4448, hsa-miR-4658, hsa-miR-22-3p, hsa-miR-4776-5p, hsa-miR-320e, hsa-miR-1225-3p, hsa-miR-6875-5p, hsa-miR-4534, hsa-miR-4652-5p, hsa-miR-648, hsa-miR-4259, hsa-miR-107, and hsa-miR-650, which are ranked based the adjusted *P*-value and correspond to SEQ ID NOS: 2-100 respectively.

[0015] According to some other embodiments of the method, the miRNA biomarker set further comprises, in addition to hsa-miR-5100, one or more of the other top 50 miRNAs listed in Table

1, i.e. hsa-miR-1343-3p, hsa-miR-1290, hsa-miR-4787-3p, hsa-miR-6877-5p, hsa-miR-17-3p, hsa-miR-6765-5p, hsa-miR-1268b, hsa-miR-4258, hsa-miR-451a, hsa-miR-1228-5p, hsa-miR-8073, hsa-miR-4454, hsa-miR-187-5p, hsa-miR-4286, hsa-miR-6746-5p, hsa-miR-663b, hsa-miR-6075, hsa-miR-5001-5p, hsa-miR-6789-5p, hsa-miR-4513, hsa-miR-3192-5p, hsa-miR-8060, hsa-miR-668-5p, hsa-miR-1268a, hsa-miR-1273g-3p, hsa-miR-4706, hsa-miR-124-3p, hsa-miR-1260b, hsa-miR-4740-5p, hsa-miR-320b, hsa-miR-7977, hsa-miR-29b-3p, hsa-miR-4708-3p, hsa-miR-4525, hsa-miR-92b-3p, hsa-miR-4257, hsa-miR-4727-3p, hsa-miR-92a-3p, hsa-miR-663a, hsa-miR-6787-5p, hsa-miR-3131, hsa-miR-6802-5p, hsa-miR-654-5p, hsa-miR-6511b-5p, hsa-miR-29b-1-5p, hsa-miR-4417, hsa-miR-4736, hsa-miR-6840-3p, and hsa-miR-4710, which are ranked based the adjusted P value and correspond to SEQ ID NOS: 2-50 respectively.

[0016] According to some other embodiments of the method, the miRNA biomarker set further comprises, in addition to hsa-miR-5100, one or more of the other top 20 miRNAs listed in Table 1, i.e. hsa-miR-1343-3p, hsa-miR-1290, hsa-miR-4787-3p, hsa-miR-6877-5p, hsa-miR-17-3p, hsa-miR-6765-5p, hsa-miR-1268b, hsa-miR-4258, hsa-miR-451a, hsa-miR-1228-5p, hsa-miR-8073, hsa-miR-4454, hsa-miR-187-5p, hsa-miR-4286, hsa-miR-6746-5p, hsa-miR-663b, hsa-miR-6075, hsa-miR-5001-5p, and hsa-miR-6789-5p, which are ranked based the adjusted P value and correspond to SEQ ID NOS: 2-20 respectively. Herein further optionally, the miRNA biomarker set consists of the top 20 miRNAs listed in Table 1 (corresponding to SEQ ID NOS: 1-20 respectively).

[0017] According to some other embodiments of the method, the miRNA biomarker set further comprises, in addition to hsa-miR-5100, one or more of the other top 4 miRNAs listed in Table 1, i.e. hsa-miR-1343-3p, hsa-miR-1290, and hsa-miR-4787-3p, which are ranked based on the adjusted P value and correspond to SEQ ID NOS: 2-4 respectively. Herein further optionally, the miRNA biomarker set consists of the top 4 miRNAs listed in Table 1, i.e. hsa-miR-5100, hsa-miR-1343-3p, hsa-miR-1290, and hsa-miR-4787-3p, which correspond to SEQ ID NOS: 1-4 respectively.

[0018] The method can optionally be further configured to be capable of achieving diagnostic accuracy having a higher AUC value.

[0019] According to some embodiments, the method is configured to be capable of achieving diagnostic accuracy having an AUC value greater than approximately 0.850. Herein optionally, the cancer that can be detected can be selected from a group consisting of lung cancer, biliary tract cancer, bladder cancer, colorectal cancer, esophageal cancer, gastric cancer, glioma cancer, liver cancer, pancreatic cancer, prostate cancer, ovarian cancer, and sarcoma.

[0020] According to some embodiments, the method is configured to be capable of achieving

diagnostic accuracy having an AUC value greater than approximately 0.950. Herein optionally, the cancer that can be detected can be selected from a group consisting of lung cancer, biliary tract cancer, bladder cancer, colorectal cancer, esophageal cancer, gastric cancer, glioma cancer, liver cancer, ovarian cancer, pancreatic cancer, and prostate cancer.

- 5 **[0021]** According to some embodiments, the method is configured to be capable of achieving diagnostic accuracy having an AUC value greater than approximately 0.990. Herein optionally, the cancer that can be detected can be selected from a group consisting of lung cancer, biliary tract cancer, bladder cancer, esophageal cancer, gastric cancer, glioma cancer, and prostate cancer.

- 10 **[0022]** According to some embodiments, the method is configured to be capable of achieving a diagnostic accuracy having an AUC value greater than approximately 0.999. Herein optionally, the cancer that can be detected can be lung cancer or gastric cancer.

[0023] Depending on different practical needs, the method can optionally be configured to be capable of achieving diagnostic accuracy having different sensitivity and specificity levels.

- 15 **[0024]** According to some embodiments, the method is configured to be capable of achieving diagnostic accuracy having a sensitivity greater than approximately 68.0% while having a specificity greater than approximately 99.0%. Herein optionally, the cancer that can be detected can be selected from a group consisting of lung cancer, biliary tract cancer, bladder cancer, colorectal cancer, esophageal cancer, gastric cancer, glioma cancer, liver cancer, pancreatic cancer, prostate cancer, ovarian cancer, and sarcoma.

- 20 **[0025]** According to some embodiments, the method is configured to be capable of achieving diagnostic accuracy having a sensitivity greater than approximately 83.0% while having a specificity greater than approximately 99.0%. Herein optionally, the cancer that can be detected can be selected from a group consisting of lung cancer, biliary tract cancer, bladder cancer, colorectal cancer, esophageal cancer, gastric cancer, glioma cancer, liver cancer, pancreatic cancer, and prostate cancer.

[0026] According to some embodiments, the method is configured to be capable of achieving diagnostic accuracy having a sensitivity greater than approximately 99.0% and having a specificity greater than approximately 99.0%. Herein optionally, the cancer that can be detected can be lung cancer or gastric cancer.

- 30 **[0027]** According to some embodiments of the method, in step (2) of calculating the diagnostic index of the biological sample based on the expression profile of the miRNA biomarker set, the diagnostic index is calculated via an unweighted model.

[0028] According to some other embodiments of the method, in step (2) of calculating the diagnostic index of the biological sample based on the expression profile of the miRNA biomarker

set, the diagnostic index is calculated via a weighted model using weights from one selected from a group consisting of Linear Models for Microarray Data (limma) model, logistic regression model, linear discriminant analysis (LDA) model, conditional logistic regression model, lasso regression model, ridge regression model, random forest, support vector machine, and probit regression model. Further optionally, the diagnostic index is calculated via a weighted model using weights from the limma model.

[0029] As used herein, the terms “unweighted model” and “weighted model” are to be understood within the common definition as well appreciated by people skilled in the art. Regarding the term “unweighted model”, it refers to a situation where no weight is applied for each miRNA in the miRNA biomarker set when calculating the diagnostic index. Within the scope of the present disclosure and with reference to formula (I), the phrase “the diagnostic index is calculated via an unweighted model” can be understood to have an equal t_i (e.g. $t_i = 1$) for any miRNA in the miRNA biomarker set. Regarding the term “weighted model”, it refers to as a situation where a corresponding weight is applied for each miRNA in the miRNA biomarker set when calculating the diagnostic index. Within the scope of the present disclosure and with reference to formula (I), the phrase “the diagnostic index is calculated via a weighted model” can be understood such that for any miRNA_i in the miRNA biomarker set, not all t_i are equal (i.e. there are at least two miRNAs which have different weights).

[0030] Each of the terms “Linear Models for Microarray Data (limma) model” (Ritchie et al. 2015), “logistic regression model” (Venable and Ripley 2002), “linear discriminant analysis (LDA) model” (Venable and Ripley 2002), “conditional logistic regression model” (Venable and Ripley 2002), “lasso regression model” (Tibshirani 1996), “ridge regression model” (Hoerl and Kennard 1970), “random forest” (Ripley 1996), “support vector machine” (Ripley 1996), and “probit regression model” (Venable and Ripley 2002) is substantially a probability-modeling statistical model that models abides by the definition commonly appreciated by people skilled in the field, the details of which can be referenced by the reference included immediately behind.

[0031] In order to provide convenience, according to some embodiments, after step (2) and before step (3), the method can further include a normalization step of: obtaining a normalized diagnostic index based on the calculated diagnostic index. Correspondingly, step (3) comprises: classifying the subject as having the cancer if the normalized diagnostic index is equal to or greater than a preset cut-point; or classifying the subject as not having the cancer if otherwise.

[0032] Herein, there can be different ways for the normalization step. According to some embodiments, the normalized diagnostic index is calculated based on formula (II):

$$\text{normalized diagnostic index} = \frac{\text{diagnostic index} - \text{param}_{\text{location}}}{\text{param}_{\text{scale}}}, \quad (\text{II})$$

Herein the $\text{param}_{\text{location}}$ and $\text{param}_{\text{scale}}$ are respectively a location parameter and a scale parameter configured to allow the normalized diagnostic index to be within a range no less than a first preset value and no greater than a second preset value.

[0033] More specifically, the $\text{param}_{\text{location}}$ is substantially a location parameter configured to shift the minimum of the normalized diagnostic index to the first preset value, and $\text{param}_{\text{scale}}$ is substantially a scale parameter configured to scale the maximum of the normalized diagnostic index to the second value. Thus the first preset value and the second preset value are respectively the minimum and maximum in the range of normalized diagnostic index values that have been obtained and calculated from an existing population of subjects known to have and known not to have the cancer, with outliers excluded.

[0034] Optionally, multiple settings can be applied. For example, in the existing dataset in EXAMPLE 1 below, where the diagnostic index values are determined to have a range of 600 to 1600 excluding outliers (see), in order to shift the range to be between 0 (i.e. the first preset value) and 10 (i.e. the second present value), the $\text{param}_{\text{location}}$ and $\text{param}_{\text{scale}}$ can be respectively set to 600 and 100 so that the final normalized diagnostic index can be no less than 0 and no greater than 10. It is note that this normalization scheme was employed in the below EXAMPLE 1.

[0035] Alternatively, the $\text{param}_{\text{location}}$ and $\text{param}_{\text{scale}}$ can be respectively set to 600 and 1000, so that the final normalized diagnostic index can be set to be no less than 0 and no greater than 1. Further alternatively, the $\text{param}_{\text{location}}$ and $\text{param}_{\text{scale}}$ can be respectively set to 600 and 10, so that the final normalized diagnostic index can be set to be no less than 0 and no greater than 100. Further alternatively, the $\text{param}_{\text{location}}$ and $\text{param}_{\text{scale}}$ can be respectively set to 350 and 250, so that the final normalized diagnostic index can be set to be no less than 1 and no greater than 5.

[0036] In embodiments where the normalized diagnostic index is normalized to be between 0 and 10, the pre-set cut-point can optionally be set as 5.1 to thereby allow the method to have a specificity that is greater than approximately 0.95, or optionally can be set as 6.0 to thereby allow the method to have a specificity that is greater than approximately 0.99.

[0037] In any embodiment of the method as described above, the biological sample is a liquid biopsy sample selected from a group consisting of a blood sample, a serum sample, a plasma sample, a urine sample (Yun et al. 2012), a saliva sample (Park et al. 2009), and a spatium sample.

[0038] In any embodiment of the method as described above, in step (1) of determining an expression profile of an miRNA biomarker set consisting of at least one miRNA from the biological sample, the expression profile of the miRNA biomarker set can optionally be obtained by means of Northern Blotting, microarray analysis, RNA-sequencing, or RNA in-situ hybridization, or can optionally be obtained by means of a nucleic acid amplification procedure,

comprising reverse-transcription PCR (RT-PCR), quantitative RT-PCR (qRT-PCR), or digital RT-PCR.

5 [0039] As used herein, each of the above miRNA detection approaches is to be understood within the common definition well-appreciated by people of ordinary skills in the field. More details for implementing these approaches to determine the expression profile of the miRNA biomarker set will be provided below.

[0040] In any embodiment of the method as described above, the method optionally further comprises a step of performing an evaluation of the subject, wherein said evaluation comprises a diagnosis of the cancer or a detection of a recurrence of the cancer.

10 [0041] Herein, the phrase “diagnosis of the cancer” is referred to as the detection of the cancer in a subject previously known not to have the cancer, whereas the phrase “recurrence of the cancer” is referred to as the detection of the cancer again in a subject with the cancer who has previously been treated to remove the cancer to become cancer-free.

15 [0042] In any embodiment of the method as described above, the method optionally further comprises a step of administering to the subject a therapeutic regimen when the subject is classified as having the cancer. Herein, a variety of known therapeutic regimens can be administered in the method, which include surgery, radiotherapy, chemotherapy, hormonal therapy, targeted therapy, immunotherapy or the combination thereof. These above therapeutic regimens have been well-established for each different cancer mentioned above.

20 [0043] In any embodiment of the method as described above, the method optionally further comprises a step of performing a diagnostic procedure on the subject when the subject is classified as having the cancer. Herein the diagnostic procedure may optionally comprise physical examination, pathological examination of a biopsy from the subject, immunohistochemistry examination, or imaging examination such as x-rays, computed tomography (CT), ultrasonography, and/or magnetic resonance imaging.

[0044] In a second aspect, the present disclosure further provides a kit for detecting a cancer from a biological sample obtained from a subject, which is substantially employed for implementing the method described in the first aspect.

30 [0045] As used herein, and elsewhere in the disclosure as well, the term “kit” is referred to as a collection of articles and/or instructions. An article included in the kit can be a physical entity or a component thereof. Examples of articles that can be included in the kit as disclosed herein can include one or more nucleic acids (e.g. polynucleotides), or one or more device, apparatus or equipment (e.g. a molecular array or microarray that comprises the one or more nucleic acids). An instruction included in the kit can be a description of the specific steps to be performed (e.g. a

manual), which can be printed on a physical medium (e.g. paper, card, etc.), on a computer-readable storage medium (e.g. hard disc, compact disc or CD, flash drive, etc.), or even stored in the internet (e.g. in an accessible cloud space), etc.

5 [0046] The kit can comprise at least the following components (1) and (2) (i.e. articles and/or instructions):

[0047] Component (1): at least one nucleic acid, each capable of specifically recognizing each miRNA in an miRNA biomarker set to thereby allow an expression profile of the miRNA biomarker set to be obtained from the biological sample. Herein the miRNA biomarker set comprises hsa-miR-5100 (SEQ ID NO: 1).

10 [0048] Component (2): at least one instruction, comprising a first instruction and a second instruction. The first instruction comprises a first sub-instruction for calculating a diagnostic index of the biological sample based on the expression profile of the miRNA biomarker set, wherein the diagnostic index is calculated based on formula:

$$\text{diagnostic index} = \sum_{i=1}^n t_i * \text{miRNA}_i; \quad (I)$$

15 where n is the total number of the at least one miRNA in the miRNA biomarker set, miRNA_i is the expression level of i^{th} miRNA in the miRNA biomarker set, i is an integer greater than zero and smaller than or equal to n , and t_i is a weight for the i^{th} miRNA. The second instruction is configured for classifying the subject as having the cancer or not, wherein the subject is classified as having the cancer if the calculated diagnostic index is greater than or equal to a pre-determined threshold
20 or as not having the cancer if otherwise.

[0049] Herein, in component (1) of the kit, the at least one nucleic acid can optionally comprise a polynucleotide capable of specifically hybridizing under a stringent condition to: either (a) a polynucleotide comprising or consisting of a nucleotide sequence of SEQ ID NO: 1, a derivative thereof, a variant thereof having at least 80% sequence identity, or a fragment thereof comprising
25 15 or more consecutive nucleotides; or (b) a polynucleotide comprising or consisting of a nucleotide sequence complementary to a nucleotide sequence of SEQ ID NO: 1, a derivative thereof, a variant thereof having at least 80% sequence identity, or a fragment thereof comprising 15 or more consecutive nucleotides.

[0050] According to some embodiments of the kit, the miRNA biomarker set further comprises,
30 in addition to hsa-miR-5100, one or more of the other 99 miRNAs listed in Table 1. Correspondingly, in component (1) of the kit, the at least one nucleic acid can optionally further comprise at least one polynucleotide, each capable of specifically hybridizing under a stringent condition to: either (a) a polynucleotide comprising or consisting of a nucleotide sequence of any one of SEQ ID NOS: 2-100, a derivative thereof, a variant thereof having at least 80% sequence

identity, or a fragment thereof comprising 15 or more consecutive nucleotides; or (b) a polynucleotide comprising or consisting of a nucleotide sequence complementary to a nucleotide sequence of any one of SEQ ID NOS: 2-100, a derivative thereof, a variant thereof having at least 80% sequence identity, or a fragment thereof comprising 15 or more consecutive nucleotides.

- 5 **[0051]** According to some embodiments of the kit, the miRNA biomarker set further comprises, in addition to hsa-miR-5100, one or more of the other top 50 miRNAs listed in Table 1. Correspondingly, in component (1) of the kit, the at least one nucleic acid can optionally further comprise at least one polynucleotide, each capable of specifically hybridizing under a stringent condition to: either (a) a polynucleotide comprising or consisting of a nucleotide sequence of any
- 10 one of SEQ ID NOS: 2-50, a derivative thereof, a variant thereof having at least 80% sequence identity, or a fragment thereof comprising 15 or more consecutive nucleotides; or (b) a polynucleotide comprising or consisting of a nucleotide sequence complementary to a nucleotide sequence of any one of SEQ ID NOS: 2-50, a derivative thereof, a variant thereof having at least 80% sequence identity, or a fragment thereof comprising 15 or more consecutive nucleotides.
- 15 **[0052]** According to some embodiments of the kit, the miRNA biomarker set further comprises, in addition to hsa-miR-5100, one or more of the other top 20 miRNAs listed in Table 1. Correspondingly, in component (1) of the kit, the at least one nucleic acid can optionally further comprise at least one polynucleotide, each capable of specifically hybridizing under a stringent condition to: either (a) a polynucleotide comprising or consisting of a nucleotide sequence of any
- 20 one of SEQ ID NOS: 2-20, a derivative thereof, a variant thereof having at least 80% sequence identity, or a fragment thereof comprising 15 or more consecutive nucleotides; or (b) a polynucleotide comprising or consisting of a nucleotide sequence complementary to a nucleotide sequence of any one of SEQ ID NOS: 2-20, a derivative thereof, a variant thereof having at least 80% sequence identity, or a fragment thereof comprising 15 or more consecutive nucleotides.
- 25 **[0053]** Herein further optionally, the miRNA biomarker set consists of the top 20 miRNAs in Table 1, and correspondingly, in component (1) of the kit, the at least one nucleic acid consists of a total of 20 polynucleotides which are respectively capable of specifically hybridizing under a stringent condition to: either (a) polynucleotides respectively comprising or consisting of nucleotide sequences of SEQ ID NOS: 1-20, derivatives thereof, variants thereof each having at
- 30 least 80% sequence identity, or fragments thereof each comprising 15 or more consecutive nucleotides; or (b) polynucleotides respectively comprising or consisting of nucleotide sequences which are respectively complementary to nucleotide sequences of SEQ ID NOS: 1-20, derivatives thereof, variants thereof each having at least 80% sequence identity, or fragments thereof each comprising 15 or more consecutive nucleotides.

[0054] According to some embodiments of the kit, the miRNA biomarker set further comprises, in addition to hsa-miR-5100, one or more of the other top 4 miRNAs listed in Table 1. Correspondingly, in component (1) of the kit, the at least one nucleic acid can optionally further comprise at least one polynucleotide, each capable of specifically hybridizing under a stringent condition to: either (a) a polynucleotide comprising or consisting of a nucleotide sequence of any one of SEQ ID NOS: 2-4, a derivative thereof, a variant thereof having at least 80% sequence identity, or a fragment thereof comprising 15 or more consecutive nucleotides; or (b) a polynucleotide comprising or consisting of a nucleotide sequence complementary to a nucleotide sequence of any one of SEQ ID NOS: 2-4, a derivative thereof, a variant thereof having at least 80% sequence identity, or a fragment thereof comprising 15 or more consecutive nucleotides.

[0055] Herein further optionally, the miRNA biomarker set consists of the top 4 miRNAs in Table 1, i.e. hsa-miR-5100, hsa-miR-1343-3p, hsa-miR-1290, and hsa-miR-4787-3p, and correspondingly, in component (1) of the kit, the at least one nucleic acid consists of a total of 4 polynucleotides which are respectively capable of specifically hybridizing under a stringent condition to: either (a) polynucleotides respectively comprising or consisting of nucleotide sequences of SEQ ID NOS: 1-4, derivatives thereof, variants thereof each having at least 80% sequence identity, or fragments thereof each comprising 15 or more consecutive nucleotides; or (b) polynucleotides respectively comprising or consisting of nucleotide sequences which are respectively complementary to nucleotide sequences of SEQ ID NOS: 1-4, derivatives thereof, variants thereof each having at least 80% sequence identity, or fragments thereof each comprising 15 or more consecutive nucleotides.

[0056] In the kit, in the first sub-instruction of the first instruction in component (2), the diagnostic index can be calculated via an unweighted model, or alternatively via a weighted model using weights from one of the probability-modeling statistical models that have been provided above in the first aspect. Herein according to some embodiments of the kit, the diagnostic index is calculated via a weighted model using weights from the limma model.

[0057] According to some embodiments of the kit, the pre-determined threshold can be set as 1110, and the second instruction further indicates that the classification using 1110 as the pre-determined threshold has a specificity > 0.95. According to some other embodiments of the kit the pre-determined threshold can be set as 1200, and the second instruction further indicates that such classification using 1200 as the pre-determined threshold has a specificity > 0.99.

[0058] According to some embodiments of the kit, the first instruction further comprises a second sub-instruction for obtaining a normalized diagnostic index based on the diagnostic index calculated according to the first sub-instruction, and in the second instruction, the subject is

classified as having the cancer if the normalized diagnostic index is greater than or equal to a preset cut-point or as not having the cancer if otherwise. The normalization process is substantially identical to the normalization process mentioned above in the first method aspect above, whose description will be skipped in here.

5 **[0059]** Optionally, the normalized diagnostic index is calculated via a weighted model using weights from the limma model, and the first preset value is 0, and the second preset value is 10. Furthermore, the preset cut-point can be set optionally as 5.1 or 6.0, to thereby allow the classification using the preset cut-point to have a specificity that is > 0.95 or > 0.99 , respectively.

10 **[0060]** According to different embodiments, the at least one instruction in component (2) in the kit may further comprise a third instruction for performing an evaluation of the subject, wherein said evaluation comprises a diagnosis of the cancer or a detection of a recurrence of the cancer; or may further comprise a fourth instruction for administering to the subject a therapeutic regimen when the subject is classified as having the cancer.

15 **[0061]** According to some embodiments, the at least one instruction in component (2) in the kit may further comprise a first additional instruction for obtaining the expression profile of the miRNA biomarker set, comprising a procedure for performing Northern Blotting, microarray analysis, RNA-sequencing, or RNA in-situ hybridization by means of the at least one nucleic acid. Herein, the at least one nucleic acid may optionally be arranged on a molecular array.

20 **[0062]** According to some embodiments, the kit may further comprise at least one set of amplification primers, each set capable of specifically amplifying each of the at least one miRNA in the miRNA biomarker set from the biological sample. As such, the at least one instruction in component (2) in the kit may further comprise a second additional instruction for obtaining the expression profile of the miRNA biomarker set, comprising a procedure for performing reverse-transcription PCR (RT-PCR), quantitative RT-PCR (qRT-PCR), or digital RT-PCR by means of the
25 at least one nucleic acid and the at least one set of amplification primers.

[0063] In any embodiment of the kit as described above, the biological sample can be a liquid biopsy sample selected from a group consisting of a blood sample, a serum sample, a plasma sample, a urine sample, a saliva sample, and a sputum sample.

30 **[0064]** In a third aspect, the present disclosure further provides a system for detecting a cancer in a subject. Herein, the system is substantially a computerized system comprising a collection of hardware (e.g. processor, memory, I/O interface, storage medium, etc.) and software (i.e. computer programs, including operation system software, and specific program software, etc.), which are configured to collaboratively work so as to collectively implement all or some steps of the method as described above in the first aspect. According to some embodiments, the system comprises a

processor and a non-transitory storage medium. The non-transitory storage medium is configured to contain a software (i.e. program instructions) for execution by the processor, and the program instructions are configured to cause the processor to execute the various steps of the method according to the various different embodiments of the method that are described above in the first aspect.

[0065] In a fourth aspect, the present disclosure further provides a non-transitory storage medium, configured to store computer-executable program instructions which, when executed by a processor, cause the processor to execute the method according to the various different embodiments of the method that are described above in the first aspect.

[0066] There can be various different embodiments for the above-mentioned system and non-transitory storage medium regarding to the following elements/features, including: what miRNA components are included in the miRNA biomarker set; whether and how a normalization is performed over the diagnostic index; how the subject is classified as having the cancer or not, what samples can be used for the biological sample, and what detection accuracy level is to be achieved, etc. The specific details for these different embodiments can be referenced to the various embodiments of the method as described in the first aspect, and will be skipped herein for conciseness.

[0067] Unless defined elsewhere, the terms as used throughout the disclosure are defined as follows.

[0068] In general terms, a “subject” means a mammal such as a primate including a human and a chimpanzee, a pet animal including a dog and a cat, a livestock animal including cattle, a horse, sheep, and a goat, and a rodent including a mouse and a rat. The term “healthy subject” also means such a mammal without the cancer to be detected. It is to be noted that the whole disclosure concerns more specifically human subjects, but can optionally be applied to other non-human mammals as well.

[0069] Unless indicated or defined otherwise, the terms or abbreviations such as “nucleic acid”, “nucleotide”, “polynucleotide”, “DNA”, “RNA”, and “miRNA” abide by common use in the art.

[0070] As used herein, the term “polynucleotide” is interchangeable with “nucleic acid”, and is referred to as a nucleic acid including all of RNA, DNA, and RNA/DNA (chimera). The DNA includes all of cDNA, genomic DNA, and synthetic DNA. The RNA includes all of total RNA, mRNA, rRNA, miRNA, siRNA, snoRNA, snRNA, non-coding RNA and synthetic RNA.

[0071] As used herein, the term “fragment” is a polynucleotide having a nucleotide sequence having a consecutive portion of a polynucleotide and desirably has a length of 15 or more nucleotides, e.g. 15, 16, 17, 18, 19, etc. nucleotides.

- [0072]** As used herein, the term “gene” is intended to include not only RNA and double-stranded DNA but also each single-stranded DNA such as a plus strand (or a sense strand) or a complementary strand (or an antisense strand) constituting the duplex. The gene is not particularly limited by its length. As used herein, the “gene” includes all of double-stranded DNA including human genomic DNA, single-stranded DNA (plus strand) including cDNA, single-stranded DNA having a sequence complementary to the plus strand (complementary strand), miRNA (miRNA), and their fragments, and their transcripts, unless otherwise specified. The “gene” includes not only a “gene” represented by a particular nucleotide sequence (or SEQ ID NO) but “nucleic acids” encoding RNAs having biological functions equivalent to an RNA encoded by the gene, for example, a congener (i.e., a homolog or an ortholog), a variant (e.g., a genetic polymorph), and a derivative. Specific examples of such a “nucleic acid” encoding a congener, a variant, or a derivative can include a “nucleic acid” having a nucleotide sequence hybridizing under stringent conditions described later to a complementary sequence of a nucleotide sequence represented by any of SEQ ID NOS: 1 to 100 or a nucleotide sequence derived from the nucleotide sequence by the replacement of the nucleotide "U" (or "u") with the nucleotide "T" (or "t"). The “gene” is not particularly limited by its functional region and can contain, for example, an expression control region, a coding region, an exon, or an intron. The “gene” may be contained in a cell or may exist alone after being released into the outside of a cell. Alternatively, the “gene” may be in a state enclosed in a vesicle called exosome.
- [0073]** Within the scope of the whole disclosure, the term “microRNA (miRNA)” is intended to mean a 15- to 25-nucleotide non-coding RNA that is transcribed as an RNA precursor having a hairpin-like structure, cleaved by a dsRNA-cleaving enzyme which has RNase III cleavage activity, integrated into a protein complex called RISC, and involved in the suppression of translation of mRNA, unless otherwise specified. The term “miRNA” as used herein includes not only a “miRNA” represented by a particular nucleotide sequence (or SEQ ID NO) but a precursor of the “miRNA” (pre-miRNA or pri-miRNA), and miRNAs having biological functions equivalent thereto, for example, a congener (i.e., a homolog or an ortholog), a variant (e.g., a genetic polymorph), and a derivative. Such a precursor, a congener, a variant, or a derivative can be specifically identified using miRBase Release 20 (Kozomara and Griffiths-Jones, 2010), and examples thereof can include an “miRNA” having a nucleotide sequence hybridizing under stringent conditions described later to a complementary sequence of any particular nucleotide sequence represented by any of SEQ ID NOS: 1 to 100. The term “miRNA” as used herein may be a gene product of a miRNA gene. Such a gene product includes a mature miRNA (e.g., a 15- to 25-nucleotide or 19- to 25-nucleotide non-coding RNA involved in the suppression of translation

of mRNA as described above) or a miRNA precursor (e.g., pre-miRNA or pri-miRNA).

[0074] As used herein, the term “probe” includes a polynucleotide that is used for specifically detecting an RNA resulting from the expression of a gene or a polynucleotide derived from the RNA, and/or a polynucleotide complementary thereto.

5 **[0075]** As used herein, the term “primer”, or “amplification primers” includes a polynucleotide that specifically recognizes and amplifies an RNA resulting from the expression of a gene or a polynucleotide derived from the RNA, and/or a polynucleotide complementary thereto.

[0076] In this context, the complementary polynucleotide (complementary strand or reverse strand) means a polynucleotide in a complementary base relationship based on A:T (U) and G:C
10 base pairs with the full-length sequence of a polynucleotide consisting of a nucleotide sequence defined by any of SEQ ID NOs: 1 to 100 or a nucleotide sequence derived from the nucleotide sequence by the replacement of the nucleotide "U" (or "u") with the nucleotide "T" (or "t"), or a partial sequence thereof (here, this full-length or partial sequence is referred to as a plus strand for the sake of convenience). However, such a complementary strand is not limited to a sequence
15 completely complementary to the nucleotide sequence of the target plus strand and may have a complementary relationship to an extent that permits hybridization under stringent conditions to the target plus strand.

[0077] As used herein, the term “stringent conditions” refers to conditions under which a nucleic acid probe hybridizes to its target sequence to a larger extent (e.g., a measurement value
20 equal to or larger than a mean of background measurement values + a standard deviation of the background measurement values×2) than that for other sequences. The stringent conditions are dependent on a sequence and differ depending on an environment where hybridization is performed. A target sequence complementary 100% to the nucleic acid probe can be identified by controlling the stringency of hybridization and/or washing conditions. Specific examples of the
25 “stringent conditions” will be mentioned later.

[0078] As used herein, the term “variant” means, in the case of a nucleic acid, a natural variant attributed to polymorphism, mutation, or the like; a variant containing the deletion, substitution, addition, or insertion of 1, 2, or 3 or more nucleotides in a nucleotide sequence represented by any of SEQ ID NOs: 1 to 100 or a nucleotide sequence derived from the nucleotide sequence by the
30 replacement of the nucleotide "U" (or "u") with the nucleotide "T" (or "t"), or a partial sequence thereof; a variant containing the deletion, substitution, addition, or insertion of 1 or 2 or more nucleotides in a nucleotide sequence of a premature miRNA of a sequence represented by any of SEQ ID NOs: 1 to 100 or a nucleotide sequence derived from the nucleotide sequence by the replacement of the nucleotide "U" (or "u") with the nucleotide "T" (or "t"), or a partial sequence

thereof; a variant that exhibits % identity of approximately 90% or higher, approximately 95% or higher, approximately 97% or higher, approximately 98% or higher, approximately 99% or higher to each of these nucleotide sequences or the partial sequences thereof; or a nucleic acid hybridizing under the stringent conditions defined above to a polynucleotide or an oligonucleotide comprising each of these nucleotide sequences or the partial sequences thereof. A variant can be prepared by use of a well-known technique such as site-directed mutagenesis or PCR-based mutagenesis.

[0079] The term “percent(%) identity” can be determined with or without an introduced gap, using a protein or gene search system based on BLAST or FASTA described above (Zhang et al., 2000; Altschul et al. 1990; Pearson et al. 1988).

[0080] The term “derivative” is meant to include a modified nucleic acid, for example, a derivative labeled with a fluorophore or the like, a derivative containing a modified nucleotide (e.g., a nucleotide containing a group such as halogen, alkyl such as methyl, alkoxy such as methoxy, thio, or carboxymethyl, and a nucleotide that has undergone base rearrangement, double bond saturation, deamination, replacement of an oxygen molecule with a sulfur atom, etc.), PNA (peptide nucleic acid; Nielsen et al. 1991), and LNA (locked nucleic acid; Obika et al. 1998) without any limitation.

[0081] The “nucleic acid” capable of specifically binding to a polynucleotide selected from the miRNAs described above is a synthesized or prepared nucleic acid and specifically includes a “nucleic acid probe” or a “primer”. The “nucleic acid” is utilized directly or indirectly for detecting the presence or absence of cancer in a subject, for diagnosing the severity, the degree of amelioration, or the therapeutic sensitivity of cancer, or for screening for a candidate substance useful in the prevention, amelioration, or treatment of cancer. The “nucleic acid” includes a nucleotide, an oligonucleotide, and a polynucleotide capable of specifically recognizing and binding to a transcript represented by any of SEQ ID NOs: 1 to 100, or a synthetic cDNA nucleic acid thereof in vivo, particularly, in a sample such as a body fluid (e.g., blood or urine), in relation to the development of cancer. The nucleotide, the oligonucleotide, and the polynucleotide can be effectively used as probes for detecting the aforementioned gene expressed in vivo, in tissues, in cells, or the like on the basis of the properties described above, or as primers for amplifying the aforementioned gene expressed in vivo.

[0082] The term “detection” as used herein is interchangeable with the term “examination”, “measurement”, or “detection or decision support”. As used herein, the term “evaluation” is meant to include diagnosis or evaluation support on the basis of examination results or measurement results.

[0083] As used within the scope of the disclosure, each of the terms “*P*-value”, “accuracy”,

“AUC”, “sensitivity”, and “specificity” is generally to be understood to have the common definition that is well appreciated by people skilled in the art, and is specifically defined as follows:

[0084] The term “*P*-value” or “*P*”, is considered to be exchangeable with “*p*-value” or “*p*”, and refers to a probability at which a more extreme statistic than that actually calculated from data under a null hypothesis is observed in a statistical test. Thus, smaller “*P*” or “*P* value” means more significant difference between subjects to be compared.

[0085] The term “AUC” means area under the curve of a Receiver Operating Characteristic curve. The term “accuracy” means a value of (the number of true positives + the number of true negatives)/(the total number of cases). The accuracy indicates the ratio of samples that were correctly identified to all samples and serves as a primary index to evaluate detection performance.

[0086] As used herein, the term “sensitivity” means a value of (the number of true positives)/(the number of true positives + the number of false negatives). High sensitivity allows cancer to be detected, leading to clinical treatment interventions.

[0087] As used herein, the term “specificity” means a value of (the number of true negatives)/(the number of true negatives + the number of false positives). High specificity prevents needless extra examination for healthy subjects misjudged as being cancer patients, leading to reduction in burden on patients and reduction in medical expense.

[0088] Unless specified elsewhere, the following summarizes the available technologies that can be used for the determination of the expression profile of the miRNA biomarker set.

[0089] It is to be noted that determination of the expression profile of the miRNA biomarker set substantially includes the determination of the expression level of each and every miRNA contained in the miRNA biomarker set. Preferably, expression levels for all of the miRNA contained in the miRNA biomarker set can be determined simultaneously in one single experiment that is well-controlled. Yet optionally, it is possible that expression levels of these miRNAs are determined in more than one experiment and by different experiment procedure.

[0090] As used herein, measuring or detecting the expression of any of the miRNAs contained in the miRNA biomarker set comprises measuring or detecting any nucleic acid transcript corresponding to the miRNA.

[0091] Typically, expression can be detected or measured on the basis of miRNA or corresponding reverse transcribed cDNA levels. Any quantitative or qualitative method for measuring RNA levels, or cDNA levels can be used. Suitable methods of detecting or measuring miRNA or cDNA levels include, for example, Northern Blotting, microarray analysis, RNA-sequencing, RNA in-situ hybridization, or a nucleic acid amplification procedure, such as reverse-transcription PCR (RT-PCR) or real-time RT-PCR, also known as quantitative RT-PCR (qRT-PCR),

or digital RT-PCR. Such methods are well known in the art (see e.g., Green and Sambrook et al. 2012). Other techniques include digital, multiplexed analysis of gene expression, such as the nCounter® (NanoString Technologies, Seattle, WA) gene expression assays, which are further described in US20100112710 and US20100047924.

- 5 **[0092]** Detecting a nucleic acid of interest generally involves hybridization between a target (e.g. miRNA or cDNA) and a probe. Sequences of the miRNAs used in various cancer gene expression profiles are known. Therefore, one of skills in the art can readily design hybridization probes for detecting those miRNAs (see e.g., Green and Sambrook et al. 2012). For example, polynucleotide probes that specifically bind to the miRNA transcripts described herein (or cDNA
- 10 synthesized therefrom) can be created using the nucleic acid sequences of the miRNA or cDNA targets themselves by routine techniques (e.g., PCR or synthesis). As used herein, the term “probe” means a part or portion of a polynucleotide sequence comprising about 10 or more contiguous nucleotides, about 15 or more contiguous nucleotides, about 20 or more contiguous nucleotides. In certain embodiments, the polynucleotide probes will comprise 10 or more nucleic acids, 15 or
- 15 more nucleic acids, or 20 or more nucleic acids. In order to confer sufficient specificity, the probe may have a sequence identity to a complement of the target sequence of about 90% or more, such as about 95% or more (e.g., about 98% or more or about 99% or more) as determined, for example, using the well-known Basic Local Alignment Search Tool (BLAST) algorithm (available through the National Center for Biotechnology Information (NCBI), Bethesda, Md.).
- 20 **[0093]** Each probe may be substantially specific for its target, to avoid any cross hybridization and false positives. An alternative to using specific probes is to use specific reagents when deriving materials from transcripts (e.g., during cDNA production, or using target-specific primers during amplification). In both cases specificity can be achieved by hybridization to portions of the targets that are substantially unique within the group of miRNAs being analyzed, for example
- 25 hybridization to the polyA tail would not provide specificity. If a target has multiple splice variants, it is possible to design a hybridization reagent that recognizes a region common to each variant and/or to use more than one reagent, each of which may recognize one or more variants.

- [0094]** Stringency of hybridization reactions is readily determinable by one of ordinary skill in the art, and generally is an empirical calculation dependent upon probe length, washing
- 30 temperature, and salt concentration. In general, longer probes may require higher temperatures for proper annealing, while shorter probes may require lower temperatures. Hybridization generally depends on the ability of denatured nucleic acid sequences to reanneal when complementary strands are present in an environment below their melting temperature. The higher the degree of desired homology between the probe and hybridizable sequence, the higher the relative

temperature that can be used. As a result, it follows that higher relative temperatures would tend to make the reaction conditions more stringent, while lower temperatures less so.

[0095] “Stringent conditions” or “high stringency conditions,” as defined herein, are identified by, but not limited to, those that: (1) use low ionic strength and high temperature for washing, for example 0.015 M sodium chloride/0.0015 M sodium citrate/0.1% sodium dodecyl sulfate at 50°C; (2) use during hybridization a denaturing agent, such as formamide, for example, 50% (v/v) formamide with 0.1% bovine serum albumin/0.1% Ficoll/0.1% polyvinylpyrrolidone/50 mM sodium phosphate buffer at pH 6.5 with 750 mM sodium chloride, 75 mM sodium citrate at 42°C; or (3) use 50% formamide, 5×SSC (0.75 M NaCl, 0.075 M sodium citrate), 50 mM sodium phosphate (pH 6.8), 0.1% sodium pyrophosphate, 5× Denhardt's solution, sonicated salmon sperm DNA (50µg/ml), 0.1% SDS, and 10% dextran sulfate at 42°C, with washes at 42°C in 0.2×SSC (sodium chloride/sodium citrate) and 50% formamide at 55°C, followed by a high-stringency wash of 0.1×SSC containing EDTA at 55°C. “Moderately stringent conditions” are described by, but not limited to, those in Sambrook et al. 1989, and include the use of washing solution and hybridization conditions (e.g., temperature, ionic strength and % SDS) less stringent than those described above. An example of moderately stringent conditions is overnight incubation at 37°C in a solution comprising: 20% formamide, 5×SSC (150 mM NaCl, 15 mM trisodium citrate), 50 mM sodium phosphate (pH 7.6), 5× Denhardt's solution, 10% dextran sulfate, and 20 mg/mL denatured sheared salmon sperm DNA, followed by washing the filters in 1×SSC at about 37-50°C. The skilled artisan will recognize how to adjust the temperature, ionic strength, etc. as necessary to accommodate factors such as probe length and the like.

[0096] In certain embodiments, microarray analysis, Northern blot, RNA in-situ hybridization, or a PCR-based method is used. In this respect, measuring the expression of the foregoing miRNAs in a biological sample can comprise, for instance, contacting a sample containing or suspected of containing cancer cells with polynucleotide probes specific to the miRNAs of interest, or with primers designed to amplify a portion of the miRNAs of interest, and detecting binding of the probes to the nucleic acid targets or amplification of the nucleic acids, respectively. Detailed protocols for designing PCR primers are known in the art (see e.g., Green and Sambrook et al. 2012). In certain embodiments, miRNAs obtained from a sample may be subjected to qRT-PCR. Reverse transcription may occur by any methods known in the art, such as through the use of an Omniscript RT Kit (Qiagen). The resultant cDNA may then be amplified by any amplification technique known in the art. miRNA expression may then be analyzed through the use of, for example, control samples as described below. As described herein, the over- or under-expression of miRNAs relative to controls may be measured to determine a miRNA expression profile for an

individual biological sample. Similarly, detailed protocols for preparing and using microarrays to analyze miRNA expression are known in the art and described herein.

[0097] As used herein, RNA-sequencing (RNA-seq), also called Whole Transcriptome Shotgun Sequencing, refers to any of a variety of high-throughput sequencing techniques used to detect the presence and quantity of RNA transcripts in real time. See Wang, Z., M. Gerstein, and M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics, NAT REV GENET, 2009. 10(1): p. 57-63. RNA-seq can be used to reveal a snapshot of a sample's miRNAs from a genome at a given moment in time. In certain embodiments, miRNA is converted to cDNA fragments via reverse transcription prior to sequencing, and, in certain embodiments, miRNA can be directly sequenced without conversion to cDNA. Adaptors may be attached to the 5' and/or 3' ends of the miRNAs, and the miRNA or cDNA may optionally be amplified, for example by PCR. The fragments are then sequenced using high-throughput sequencing technology, such as, for example, those available from Roche (e.g., the 454 platform), Illumina, Inc., and Applied Biosystem (e.g., the SOLiD system).

BRIEF DESCRIPTION OF THE DRAWINGS

[0098] FIGS. 1A-1C show a case flow diagram for lung cancer dataset (FIG. 1A, split into a discovery and a validation set) and for ovarian, liver and bladder cancer datasets (FIG. 1B, combined into a single validation dataset after removing redundant samples), and summarize the patient and tumor characteristics of patients with lung, bladder, ovarian, and liver cancers and demographic information of the corresponding controls (FIG. 1C);

[0099] FIGS. 2A-2G show the development and validation of the 4-miRNA diagnostic model in the lung cancer data set, with FIG. 2A showing determination of the optimal number (dotted line) of miRNAs for the diagnostic model by 10-fold cross validation in the discovery set; FIG. 2B showing ROC analysis in the discovery set; FIG. 2C showing distribution of normalized diagnostic index in the discovery set; FIG. 2D showing ROC analysis in the validation set; FIG. 2E showing distribution of normalized diagnostic index in the validation set; FIG. 2F showing comparison of normalized diagnostic index of paired serum samples (pre- vs. post-surgery) of 180 lung cancer patients; and FIG. 2G showing distribution of normalized diagnostic index in the clinical subsets of the validation set. Dotted horizontal lines represent the cut-point for the normalized diagnostic index of our model. The percentages shown in the graph were sensitivities in each cancer subgroup.

[0100] FIGS. 3A and 3B show the performance of 4-miRNA diagnostic model in the datasets of additional cancers, with FIG. 3A showing ROC analysis, and FIG. 3B showing distribution of

normalized diagnostic index the 4-miRNA model. The percentages shown in the graph were sensitivities of each cancer type and specificity of non-cancer controls;

[0101] FIGS. 4A and 4B show the ROC analysis and distribution of normalized diagnostic index across age and gender groups in the lung cancer dataset.

5

DETAILED DESCRIPTION

[0102] The present disclosure provides an approach, comprising a method, a kit and a computerized system, that is capable of accurately and reliably detecting one or multiple human cancers for a subject based on the expression profile of an miRNA biomarker set consisting of at least one miRNA that is determined from a biological sample obtained from the subject.

[0103] In the first aspect of this section, a detection method capable of achieving diagnostic accuracy having an AUC value greater than approximately 0.780 is provided, which substantially includes the following three steps:

[0104] Step (1): determining the expression profile of the miRNA biomarker set;
 15 [0105] Step (2): calculating a diagnostic index of the biological sample based on the expression profile of the miRNA biomarker set. The diagnostic index is calculated based on:

$$\text{diagnostic index} = \sum_{i=1}^n t_i * \text{miRNA}_i; \quad (I)$$

where n is the total number of miRNAs in the miRNA biomarker set, miRNA_i is the expression level of the i^{th} miRNA in the miRNA biomarker set, where i is an integer greater than zero and smaller than or equal to n ; and t_i is a weight for the i^{th} miRNA; and

[0106] Step (3): classifying the subject as having the cancer or not based on the value of the calculated diagnostic index. If the calculated diagnostic index is greater than or equal to a pre-determined threshold, the subject is classified as having the cancer; or if otherwise the subject is classified as not having the cancer.

25 [0107] Herein, the miRNA biomarker set includes hsa-miR-5100, and optionally can further include any one or a combination of the miRNAs listed in Table 1 (see EXAMPLE 1). According to different embodiments, in addition to hsa-miR-5100, the miRNA biomarker set may further include miRNA(s) from the top 2-100 miRNAs, or alternatively may further include miRNA(s) from the top 2-50 miRNAs, or alternatively may further include miRNA(s) from the top 2-20 miRNAs, or alternatively may further include miRNA(s) from the top 2-4 miRNAs, in Table 1.

30 [0108] Preferably, the miRNA biomarker set consists of the top 4 miRNAs (i.e. hsa-miR-5100, hsa-miR-1343-3p, hsa-miR-1290, and hsa-miR-4787-3p). Herein, depending on different embodiments, there can be different AUC cut-off levels (e.g. 0.780, 0.850, 0.950, 0.990, and 0.999), or different sensitivity-specificity levels (e.g. 68%-99%, 68%-99%, 83%-99%, and 99%-99%), at

least at which the method is capable of accurately detecting certain cancer types. For example, the method can accurately detect lung cancer and gastric cancer at the AUC > 0.999, and/or at a sensitivity > 99.0% and having a specificity > 99.0%.

[0109] There can be different ways to calculate the diagnostic index based on formula (I).

5 Optionally, the calculation can be based on an unweighted model or on a weighted model. Under the latter situation, different models (e.g. limma model, logistic regression model, etc.) may optionally be applied for obtaining the weights for the miRNAs in the miRNA biomarker set.

[0110] Preferably, the diagnostic index is calculated via a weighted model using weights from the limma model. Herein, in step (3) of the method, the pre-determined threshold can be set as

10 1110 to thereby allow the method to have a specificity >0.95; or optionally, the pre-determined threshold can be set as 1200 to thereby allow the method to have a specificity >0.99.

[0111] Optionally the diagnostic index calculated in step (2) can further undergo a normalization process, and the step (3) can determine the cancer classification based on whether the normalized diagnostic index is no less than or greater than a preset cut-point.

15 [0112] It is noted that selection of the normalization process is arbitrary. According to some embodiments, the normalization process can be based on formula:

$$\text{normalized diagnostic index} = \frac{\text{diagnostic index} - \text{param}_{\text{location}}}{\text{param}_{\text{scale}}}; \quad (\text{II})$$

where the $\text{param}_{\text{location}}$ and $\text{param}_{\text{scale}}$ are respectively a location parameter and a scale parameter configured to allow the normalized diagnostic index to be within a range no less than a first

20 preset value and no greater than a second preset value.

[0113] Herein, optionally, the $\text{param}_{\text{location}}$ and $\text{param}_{\text{scale}}$ can be selected as 600 and 1000 respectively to thereby allow the normalized diagnostic index to be between 0 and 10, and under such normalization, the preset cut-point can be set as 5.1 to give a specificity > 0.95 or as 6.0 to give a specificity > 0.99.

25 [0114] In the method, the biological sample can advantageously be a liquid biopsy sample such as a blood sample, a serum sample, a plasma sample, a urine sample, a saliva sample, or a spatium sample, etc. Determination of the expression profile of the miRNA biomarker set can be realized by means of a variety of probe-based approaches including Northern Blotting, microarray analysis, RNA-sequencing, or RNA in-situ hybridization, or by means of a variety of
30 amplification-dependent approaches including reverse-transcription PCR (RT-PCR), quantitative RT-PCR (qRT-PCR), or digital RT-PCR.

[0115] Optionally, the method may further comprise a step of performing an evaluation of the subject, so as to determine if the subject is diagnosed as having the cancer (if the subject is absent of cancer before) or if the subject has recurrence of the cancer (if the subject has been treated to

remove, or be free of, the cancer before). For such a purpose, the evaluation may further include physical examination, pathological examination of a biopsy from the subject, immunohistochemistry examination, or imaging examination including x-rays, computed tomography (CT), ultrasonography, magnetic resonance imaging, etc.

- 5 **[0116]** Further optionally, the method may further comprise a step of administering to the subject a therapeutic regimen, such as surgery, radiotherapy, chemotherapy, hormonal therapy, targeted therapy, immunotherapy or the combination thereof, when the subject is classified as having the cancer.

- 10 **[0117]** In the second aspect, a kit that can be employed to specifically implement the various steps of the method according to the different embodiments as described above in the first aspect of this section is further provided.

- 15 **[0118]** The kit substantially include certain articles (i.e. component (1), including one or more nucleic acids that can specifically recognize each miRNA in the miRNA biomarker set, and optionally one or more amplification primers) that can be used to determine the expression profile of the miRNA biomarker set and certain instructions (i.e. component (2)) for calculating the diagnostic index and for cancer classification.

- 20 **[0119]** Depending on the miRNAs included in the miRNA biomarker set, each of the nucleic acids in component (1) may comprise a polynucleotide capable of specifically hybridizing under a stringent condition to (a) a polynucleotide comprising or consisting of a nucleotide sequence as set forth in SEQ ID NOS: 1-100, 1-50, 1-20 or 1-4, a derivative thereof, a variant thereof having at least 80% sequence identity, or a fragment thereof comprising 15 or more consecutive nucleotides; or (b) a polynucleotide comprising or consisting of a nucleotide sequence complementary to a nucleotide sequence of SEQ ID NOS: 1-100, 1-50, 1-20 or 1-4, a derivative thereof, a variant thereof having at least 80% sequence identity, or a fragment thereof comprising 25 15 or more consecutive nucleotides.

- 30 **[0120]** There can be various different embodiments for the kit regarding to the following elements/features, including: what miRNA components are included in the miRNA biomarker set; whether and how a normalization is performed over the diagnostic index; how the subject is classified as having the cancer or not, what samples can be used for the biological sample, and what detection accuracy level is to be achieved, etc. The specific details for these different embodiments can be referenced to the various embodiments of the method as described above, and will be skipped herein for conciseness.

- 35 **[0121]** In the third aspect of this section, a computerized solution is further provided, which substantially serves, in a computerized and automatic manner, to implement the various steps of

the method as described above in the first aspect of this section.

[0122] Such a computerized solution may be applied in a situation where the implementation of the various steps (1)-(3) of the method described above is to be automated by running a software program comprising program instructions in a computer, which brings about advantages such as high efficiency and great convenience.

[0123] Specifically, such a computerized solution may include a computerized system or computer system, which comprises a processor (i.e. controller) and a computer-readable non-transitory storage medium that is communicatively coupled to the processor. The computer-readable non-transitory storage medium is configured to store program instructions that are executable by the processor, thereby causing the processor to execute the various different steps in the method as described above, including:

[0124] Step (1): determining the expression profile of the miRNA biomarker set;

[0125] Step (2): calculating a diagnostic index of the biological sample based on the expression profile of the miRNA biomarker set and according to formula (I); and

[0126] Step (3): classifying the subject as having the cancer or not based on the value of the calculated diagnostic index.

[0127] As used herein, the “processor” is interpreted to be exchangeable with “central controller” or “central computing unit (CPU)”, and can be deemed to be a single core or multi core processor, or a plurality of processors for parallel processing. The term “non-transitory,” as used herein, is intended to describe a tangible computer-readable storage medium excluding propagating electromagnetic signals, but are not intended to otherwise limit the type of physical computer-readable storage device that is encompassed by the phrase. Examples may include any tangible or non-transitory storage media or memory media such as electronic, magnetic, or optical media (e.g., disk or CD/DVD-ROM), or non-volatile memory storage (e.g., “flash” memory), etc.

[0128] As illustrated in FIG. 5, the system 100 can, in addition to the processor 10 and the computer-readable non-transitory storage medium 20, further comprise a bus 30, a memory 40, an I/O interface 50, and a communication interface 60. The processor 10, the storage medium 20, the memory 40, the I/O interface 50 and the communication interface 60 are all communicatively coupled with one another through the bus 30.

[0129] The storage medium 20 stores computer-executable program instructions which, when executed by the processor 10, cause the processor 10 to execute steps (1)-(3) of the method as described above. The memory 40 is configured to transiently store the program instructions obtained from the storage medium 20, and the processor 10 is configured to execute the program instructions transiently stored in the memory 40. The I/O interface 50 allows an input/output

between the system 100 and a user, realizing the control of the system 100. The communication interface 60 can allow the system 100 to be communicatively connected to another computing device to exchange data. It is to be noted that these computer hardware components can be locally arranged, or can be remotely arranged via a network, such as an intranet, an internet, or a cloud.

5 [0130] In the following, one example is provided to illustrate the inventions as described above in the various aspects of the disclosure.

[0131] EXAMPLE 1

[0132] In this example, development and validation of a circulating cell-free miRNA-based diagnostic signature for MCED is provided by utilizing four large miRNA microarray datasets, all
10 based on a standardized microarray platform.

[0133] 2. Materials and Methods

[0134] 2.1. Study Design

[0135] Four microarray datasets totaling 7536 unique participants including 3604 cancer patients and 3932 non-cancer controls were included in the current analysis, all derived from
15 studies originating from a Japanese nationwide research project “Development and Diagnostic Technology for Detection of miRNA in Body Fluids” designed to characterize serum miRNAs in over 50,000 participants across 13 cancer types using a standardized microarray platform (Asakura et al. 2020; Yokoi et al. 2018; Usuba et al. 2019, Yamamoto et al. 2020). The four datasets were originally assembled to develop diagnostic signatures for lung (GSE137140), ovarian
20 (GSE106817), liver (GSE113740), and bladder (GSE113486) cancers, respectively.

[0136] The lung cancer dataset has the largest sample size for a single cancer type (n=1566) and non-cancer controls (n=2178). The original lung cancer study established a 2-miRNA diagnostic model (referred to as the “original 2-miRNA model” in this study) with high sensitivity and specificity for the detection of lung cancer (Asakura et al. 2020). The objective of the current
25 study was initially set to use this dataset to develop and validate a new diagnostic model that may out-perform the original 2-miRNA model for lung cancer detection. As datasets for additional cancer types were identified, the new model was evaluated for performance to detect other cancers.

[0137] 2.2. Participants and Serum Samples

[0138] Serum sample collection has been previously described in the original publications
30 (Asakura et al. 2020; Yokoi et al. 2018; Usuba et al. 2019, Yamamoto et al. 2020). Briefly, serum samples were collected from cancer patients who were referred or admitted to the National Cancer Center Hospital (NCCCH) between 2008 to 2016 prior to surgical operation, and stored at 4 °C for one week before being stored at –20 °C until further use. Cancer patients who were treated with preoperative chemotherapy and radiotherapy prior to serum collection were excluded. The serum

samples for non-cancer controls who had no history of cancer and no hospitalization during the previous 3 months were collected along with routine blood tests from outpatient departments of three sources: NCCH, National Center for Geriatrics and Gerontology (NCGG) Biobank, and Yokohama Minoru Clinic (YMC). Serums collected from NCCH were stored in the same way as the cancer patients, while those from NCGG and YMC were stored at -80°C till use. The original studies were approved by the NCCH Institutional Review Board, the Ethics and Conflict of Interest Committee of the NCGG, and the Research Ethics Committee of Medical Corporation Shintokai YMC. Written informed consent was obtained from each participant.

[0139] 2.3. miRNA Microarray Expression Analysis

[0140] Details about microarray analysis were described in the original publications (Asakura et al. 2020; Yokoi et al. 2018; Usuba et al. 2019, Yamamoto et al. 2020). Briefly, total RNA was extracted from 300 μL serum, labeled by 3D-Gene[®] miRNA Labeling kit and hybridized to 3D-Gene[®] Human miRNA Oligo Chip (Toray Industries, Kanagawa, Japan) designed to investigate 2588 miRNA sequences registered in miRBase release 21. The following low-quality samples were excluded: coefficient of variation of negative control probes >0.15 ; and number of flagged probes identified by 3D-Gene[®] Scanner as “uneven spot images” >10 . The presence of a miRNA was determined when signal intensity was greater than mean plus two times standard deviation of the negative control signals, and in using the negative control signals the top and bottom 5% of the ranked signal intensities were removed. Background subtraction was performed by subtracting the mean signal of negative control signals (after removing top and bottom 5% as ranked by signal intensities) from the miRNA signal. Normalization across microarrays was achieved by calibrating according to three pre-selected internal control miRNAs (miR-149-3p, miR-2861, and miR-4463).

[0141] 2.4. Diagnostic Model Development

[0142] Patients in the lung cancer dataset were divided into the same discovery and validation sets as in the original publication (FIG. 1A) (Asakura et al. 2020), because (1) the discovery set was selected by the original authors to be balanced between cancer and non-cancer with respect to age, sex, and smoking history; (2) 50% of non-cancer participants in the discovery set were from NCCH with the same serum storage condition as cancer patients to minimize potential bias in miRNA candidates selection; (3) Using the same discovery and validation sets allows direct performance comparison of the new diagnostic model with the original 2-miRNA model. As the diagnostic model was developed from the lung cancer discovery set, after its validation in the lung cancer validation set, we further tested its ability as a multi-cancer diagnostic model in a combined dataset of other additional cancer types that were not used in the model development.

[0143] Linear Model for Microarray Data (limma) (Ritchie et al. 2015) was performed in the

discovery set to evaluate the statistical significance of differential miRNA expression between lung cancer vs. non-cancer. Ten-fold cross validation in the discovery set, based on the area under the curve (AUC) of the Receiver's Operating Characteristics (ROC) curve analysis, was performed to determine the optimal number of miRNAs for the best diagnostic model. A diagnostic index was calculated as a linear sum of miRNA expression levels weighted by limma statistics. The cut-point for the diagnostic index was chosen to ensure no misclassification of non-cancer controls in the discovery set to minimize false positives as the diagnostic model may potentially be used as a screening test in the at-risk general public.

[0144] 2.5. Statistical Analysis

The diagnostic performance for identifying cancer vs. non-cancer was determined by AUC of the ROC curve analysis, sensitivity, and specificity. Comparing AUC of two ROC curves was done with roc.test function with bootstrapping method from pROC package. Comparing paired sensitivities for the lung cancer clinical subsets of paired pre- vs. post-surgical samples was performed by McNemar test. limma analysis was carried out using Bioconductor package limma (The Bioconductor Open Source Software For Bioinformatics (accessed on August 27, 2020)). All statistical analysis was performed using R version 4.0.5 (The R Project for Statistical Computing (accessed on July 15, 2020)).

[0146] 3. Results

[0147] 3.1. Participants and Datasets

The lung cancer dataset included 1566 lung cancer patients and 2178 non-cancer controls (FIG. 1A) (Asakura et al. 2020). The ovarian cancer dataset consisted of 333 ovarian cancer patients and 2759 non-cancer controls, as well as patients with breast, colorectal, esophageal, gastric, liver, lung, pancreatic, and sarcoma cancers (FIG. 1B) (Yokoi et al. 2018). The liver and bladder cancer datasets included 345 liver cancer/1033 non-cancer and 392 bladder cancer/100 non-cancer participants, respectively, in addition to patients with biliary tract, breast, colorectal, esophageal, gastric, glioma, lung, ovarian, pancreatic, prostate, and sarcoma cancers (FIG. 1B) (Usuba et al. 2019, Yamamoto et al. 2020). With the lung cancer dataset left intact, redundant samples within the other three datasets that showed correlations either among themselves or with samples in the lung cancer dataset being greater than 0.99 were removed. Then, the unique samples from the ovarian, liver, and bladder cancer datasets were then combined into a single non-lung cancer dataset with a total of 3792 samples, including 2038 cancer patients across 12 cancer types and 1754 non-cancer controls (FIG. 1B).

The lung cancer dataset was divided into the same discovery set (n=416) and validation set (n=3328) as the original study (FIG. 1A). The discovery set included 208 lung cancer patients

and 208 non-cancer controls, matched by age, sex, and smoking status (Asakura et al. 2020). The validation set included 1358 lung cancer patients and 1970 non-cancer controls. The patients with lung cancer included 57% male, 62% former or current smokers, 78% adenocarcinoma, 14% squamous carcinoma, 72% stage I, 15% stage II, and 13% stage III (FIG. 1C).

- 5 [0150] The 392 bladder cancer patients were of mean age 68 y, 72% male, 5% metastatic, 12% nodal positive, 77% T2 or below, and 80% high grade (FIG. 1C). The 333 ovarian cancer patients were of mean age 57 y, 25% stage I, 10% stage II, 55% serous, 19% clear cell, and 13% endometrioid histology (FIG. 1C). The 348 liver cancer patients were of mean age 68 y, 78% male, 37% stage I, and 33% stage II (FIG. 1C). No detailed demographic information and tumor
- 10 characteristics for the other cancers were provided by the original studies.

Table 1. Top 100 differentially expressed miRNAs from the lung cancer discovery set.

miRNA name	miRBase Accession ID	Log Fold Change	Adjusted P-value	AUC of ROC	Sequence (SEQ ID NO.)
hsa-miR-5100	MIMAT0022259	3.931	9.99E-176	0.9988	UUCAGAUCCAGCGGUGCCUCU (SEQ ID NO. 1)
hsa-miR-1343-3p	MIMAT0019776	2.609	5.83E-94	0.9690	CUCCUGGGGCCCGCACUCUCGC (SEQ ID NO. 2)
hsa-miR-1290	MIMAT0005880	6.538	2.22E-87	0.9979	UGGAUUUUUGGAUCAGGGA (SEQ ID NO. 3)
hsa-miR-4787-3p	MIMAT0019957	1.854	6.81E-67	0.9352	GAUGCGCCGCCACUGCCC CGCGC (SEQ ID NO. 4)
hsa-miR-6877-5p	MIMAT0027654	1.364	1.37E-63	0.9490	AGGGCCGAAGGGUGGAAGCUGC (SEQ ID NO. 5)
hsa-miR-17-3p	MIMAT0000071	4.088	1.66E-62	0.9346	ACUGCAGUGAAGGCACUUGUAG (SEQ ID NO. 6)
hsa-miR-6765-5p	MIMAT0027430	-0.688	5.09E-62	0.9254	GUGAGGCGGGGCCAGGAGGGUGUGU (SEQ ID NO. 7)
hsa-miR-1268b	MIMAT0018925	-0.618	1.45E-61	0.9374	CGGGCGUGGUGGUGGGGUG (SEQ ID NO. 8)
hsa-miR-4258	MIMAT0016879	1.777	6.45E-59	0.8983	CCCCGCCACCGCCUUGG (SEQ ID NO. 9)
hsa-miR-451a	MIMAT0001631	5.008	3.71E-58	0.9384	AAACCGUUACCAUACUGAGUU (SEQ ID NO. 10)
hsa-miR-1228-5p	MIMAT0005582	-0.780	1.01E-57	0.9175	GUGGGCGGGGCAGGUGUGUG (SEQ ID NO. 11)
hsa-miR-8073	MIMAT0031000	2.087	6.42E-55	0.9058	ACCUGGCAGCAGGGAGCGUCGU (SEQ ID NO. 12)
hsa-miR-4454	MIMAT0018976	1.661	3.91E-51	0.8982	GGAUCCGAGUCACGGCACC A (SEQ ID NO. 13)
hsa-miR-187-5p	MIMAT0004561	1.044	1.90E-50	0.9126	GGCUACAACACAGGACCCG GGC (SEQ ID NO. 14)
hsa-miR-4286	MIMAT0016916	1.590	1.05E-49	0.8849	ACCCACUCCUGGUACC (SEQ ID NO. 15)
hsa-miR-6746-5p	MIMAT0027392	1.346	1.53E-49	0.8734	CCGGGAGAAGGAGGUGGC CUGG (SEQ ID NO. 16)
hsa-miR-663b	MIMAT0005867	1.201	9.31E-49	0.8872	GGUGGCCCGGCCGUGCCUG AGG (SEQ ID NO. 17)
hsa-miR-6075	MIMAT0023700	0.794	2.77E-47	0.8913	ACGGCCCAGGCGGCAUUG GUG (SEQ ID NO. 18)

hsa-miR-5001-5p	MIMAT0021021	0.796	3.16E-46	0.8841	AGGGCUGGACUCAGCGGC GGAGCU (SEQ ID NO. 19)
hsa-miR-6789-5p	MIMAT0027478	0.683	6.98E-46	0.8925	GUAGGGGCGUCCCGGGCG CGCGGG (SEQ ID NO. 20)
hsa-miR-4513	MIMAT0019050	1.063	1.19E-45	0.8946	AGACUGACGGCUGGAGGC CCAU (SEQ ID NO. 21)
hsa-miR-3192-5p	MIMAT0015076	4.111	1.76E-45	0.8605	UCUGGGAGGUUGUAGCAG UGGAA (SEQ ID NO. 22)
hsa-miR-8060	MIMAT0030987	3.502	1.77E-45	0.8779	CCAUGAAGCAGUGGGUAG GAGGAC (SEQ ID NO. 23)
hsa-miR-668-5p	MIMAT0026636	2.748	2.02E-45	0.8934	UGCGCCUCGGGUGAGCAU G (SEQ ID NO. 24)
hsa-miR-1268a	MIMAT0005922	-0.610	2.40E-45	0.8739	CGGGCGUGGUGGUGGGGG (SEQ ID NO. 25)
hsa-miR-1273g-3p	MIMAT0022742	1.448	2.67E-45	0.8620	ACCACUGCACUCCAGCCUG AG (SEQ ID NO. 26)
hsa-miR-4706	MIMAT0019806	1.063	5.43E-45	0.8509	AGCGGGGAGGAAGUGGGC GCUGCUU (SEQ ID NO. 27)
hsa-miR-124-3p	MIMAT0000422	3.734	5.43E-45	0.8903	UAAGGCACGCGUGAAUG CCAA (SEQ ID NO. 28)
hsa-miR-1260b	MIMAT0015041	1.278	9.38E-45	0.8641	AUCCCACCACUGCCACCAU (SEQ ID NO. 29)
hsa-miR-4740-5p	MIMAT0019869	3.165	9.50E-45	0.8884	AGGACUGAUCCUCUCGGG CAGG (SEQ ID NO. 30)
hsa-miR-320b	MIMAT0005792	2.317	1.08E-44	0.8868	AAAAGCUGGGUUGAGAGG GCAA (SEQ ID NO. 31)
hsa-miR-7977	MIMAT0031180	1.267	4.78E-43	0.8679	UUCCCAGCCAACGCACCA (SEQ ID NO. 32)
hsa-miR-29b-3p	MIMAT0000100	4.104	1.07E-42	0.8607	UAGCACCAUUGAAAUCA GUGUU (SEQ ID NO. 33)
hsa-miR-4708-3p	MIMAT0019810	2.780	2.73E-42	0.8571	AGCAAGGCGGCAUCUCUC UGAU (SEQ ID NO. 34)
hsa-miR-4525	MIMAT0019064	2.389	3.12E-42	0.8480	GGGGGGAUGUGCAUGCUG GUU (SEQ ID NO. 35)
hsa-miR-92b-3p	MIMAT0003218	2.494	3.43E-42	0.8677	UAUUGCACUCGUCCCGGCC UCC (SEQ ID NO. 36)
hsa-miR-4257	MIMAT0016878	1.007	4.69E-42	0.8588	CCAGAGGUGGGGACUGAG (SEQ ID NO. 37)
hsa-miR-4727-3p	MIMAT0019848	2.681	7.55E-42	0.8641	AUAGUGGGAAGCUGGCAG AUUC (SEQ ID NO. 38)
hsa-miR-92a-3p	MIMAT0000092	2.012	9.49E-42	0.8628	UAUUGCACUUGUCCCGGCC UGU (SEQ ID NO. 39)
hsa-miR-663a	MIMAT0003326	1.077	1.02E-41	0.8429	AGGCGGGGCGCCGCGGGA CCGC (SEQ ID NO. 40)
hsa-miR-6787-5p	MIMAT0027474	1.234	5.33E-41	0.8343	UGGCGGGGUAGAGCUGG CUGC (SEQ ID NO. 41)
hsa-miR-3131	MIMAT0014996	1.186	7.21E-41	0.8529	UCGAGGACUGGUGGAAGG GCCUU (SEQ ID NO. 42)
hsa-miR-6802-5p	MIMAT0027504	0.851	2.03E-40	0.8382	CUAGGUGGGGGGCUUGAA GC (SEQ ID NO. 43)
hsa-miR-654-5p	MIMAT0003330	2.538	3.90E-40	0.8698	UGGUGGGCCGCAGAACAU GUGC (SEQ ID NO. 44)
hsa-miR-6511b-5p	MIMAT0025847	1.931	1.70E-39	0.8943	CUGCAGGCAGAAGUGGGG CUGACA (SEQ ID NO. 45)

hsa-miR-29b-1-5p	MIMAT0004514	4.291	1.38E-38	0.8268	GCUGGUUUCUAUAUGGUGG UUUAGA (SEQ ID NO. 46)
hsa-miR-4417	MIMAT0018929	0.424	1.66E-38	0.8815	GGUGGGCUUCCCGGAGGG (SEQ ID NO. 47)
hsa-miR-4736	MIMAT0019862	1.509	2.07E-38	0.8664	AGGCAGGUUAUCUGGGCU G (SEQ ID NO. 48)
hsa-miR-6840-3p	MIMAT0027583	0.913	3.82E-38	0.8361	GCCCAGGACUUUGUGCGG GGUG (SEQ ID NO. 49)
hsa-miR-4710	MIMAT0019815	2.579	4.97E-38	0.8454	GGGUGAGGGCAGGUGGUU (SEQ ID NO. 50)
hsa-miR-4635	MIMAT0019692	2.883	6.15E-38	0.8521	UCUUGAAGUCAGAACCCG CAA (SEQ ID NO. 51)
hsa-miR-296-3p	MIMAT0004679	1.513	1.29E-37	0.8258	GAGGGUUGGGUGGAGGCU CUCC (SEQ ID NO. 52)
hsa-miR-1199-5p	MIMAT0031119	1.674	1.86E-37	0.9206	CCUGAGCCCGGCCGCGCA G (SEQ ID NO. 53)
hsa-miR-7975	MIMAT0031178	1.390	2.19E-37	0.8393	AUCCUAGUCACGGCACCA (SEQ ID NO. 54)
hsa-miR-4480	MIMAT0019014	3.982	4.89E-37	0.8496	AGCCAAGUGGAAGUUACU UUA (SEQ ID NO. 55)
hsa-miR-3648	MIMAT0018068	0.970	5.72E-37	0.8367	AGCCGCGGGGAUCGCCGA GGG (SEQ ID NO. 56)
hsa-miR-371a-5p	MIMAT0004687	0.870	6.10E-37	0.8597	ACUCAACUGUGGGGGCA CU (SEQ ID NO. 57)
hsa-miR-4771	MIMAT0019925	3.676	8.98E-37	0.8670	AGCAGACUUGACCUACAA UUA (SEQ ID NO. 58)
hsa-miR-6717-5p	MIMAT0025846	1.864	1.57E-36	0.8297	AGGCGAUGUGGGGAUGUA GAGA (SEQ ID NO. 59)
hsa-miR-1254	MIMAT0005905	1.180	1.74E-36	0.8502	AGCCUGGAAGCUGGAGCC UGCAGU (SEQ ID NO. 60)
hsa-miR-1246	MIMAT0005898	4.358	2.88E-36	0.8329	AAUGGAUUUUUGGAGCAG G (SEQ ID NO. 61)
hsa-miR-23b-3p	MIMAT0000418	3.275	4.55E-36	0.8473	AUCACAUUGCCAGGGAUU ACCAC (SEQ ID NO. 62)
hsa-miR-320a	MIMAT0000510	1.560	6.90E-36	0.8407	AAAAGCUGGGUUGAGAGG GCGA (SEQ ID NO. 63)
hsa-miR-4687-5p	MIMAT0019774	1.025	1.02E-35	0.8424	CAGCCCUCCUCCCGCACCC AAA (SEQ ID NO. 64)
hsa-miR-191-5p	MIMAT0000440	3.613	2.26E-35	0.8409	CAACGGAUCCCAAAGC AGCUG (SEQ ID NO. 65)
hsa-miR-320c	MIMAT0005793	2.483	2.27E-35	0.8688	AAAAGCUGGGUUGAGAGG GU (SEQ ID NO. 66)
hsa-miR-6131	MIMAT0024615	3.119	4.64E-35	0.7915	GGCUGGUCAGAUGGGAGU G (SEQ ID NO. 67)
hsa-miR-4515	MIMAT0019052	2.382	4.69E-35	0.8161	AGGACUGGACUCCCGGCA GCCC (SEQ ID NO. 68)
hsa-miR-342-5p	MIMAT0004694	1.848	4.72E-35	0.8405	AGGGGUGCUAUCUGUGAU UGA (SEQ ID NO. 69)
hsa-miR-4718	MIMAT0019831	3.469	4.73E-35	0.8508	AGCUGUACCUGAAACCAA GCA (SEQ ID NO. 70)
hsa-miR-23a-3p	MIMAT0000078	3.010	5.66E-35	0.8385	AUCACAUUGCCAGGGAUU UCC (SEQ ID NO. 71)
hsa-miR-4455	MIMAT0018977	2.615	6.45E-35	0.8547	AGGGUGUGUGUUUUU (SEQ ID NO. 72)

hsa-miR-211-3p	MIMAT0022694	1.267	1.43E-34	0.8042	GCAGGGACAGCAAAGGGG UGC (SEQ ID NO. 73)
hsa-miR-3122	MIMAT0014984	1.673	2.63E-34	0.8801	GUUGGGACAAGAGGACGG UCUU (SEQ ID NO. 74)
hsa-miR-103a-3p	MIMAT0000101	3.839	4.23E-34	0.8246	AGCAGCAUUGUACAGGGC UAUGA (SEQ ID NO. 75)
hsa-miR-4429	MIMAT0018944	1.427	1.36E-33	0.8235	AAAAGCUGGGCUGAGAGG CG (SEQ ID NO. 76)
hsa-miR-920	MIMAT0004970	2.319	1.79E-33	0.8328	GGGAGCUGUGGAAGCAG UA (SEQ ID NO. 77)
hsa-miR-3194-3p	MIMAT0019218	3.177	1.93E-33	0.8231	AGCUCUGCUGCUCACUGGC AGU (SEQ ID NO. 78)
hsa-miR-4754	MIMAT0019894	3.293	2.21E-33	0.8155	AUGCGGACCUGGGUAGC GGAGU (SEQ ID NO. 79)
hsa-miR-1238-5p	MIMAT0022947	1.377	2.33E-33	0.7838	GUGAGUGGGAGCCCCAGU GUGUG (SEQ ID NO. 80)
hsa-miR-3191-3p	MIMAT0015075	1.600	2.38E-33	0.8673	UGGGGACGUAGCUGGCCA GACAG (SEQ ID NO. 81)
hsa-miR-4755-3p	MIMAT0019896	3.798	3.45E-33	0.8298	AGCCAGGCUCUGAAGGGA AAGU (SEQ ID NO. 82)
hsa-miR-3688-5p	MIMAT0019223	3.645	7.47E-33	0.8113	AGUGGCAAAGUCUUUCCA UAU (SEQ ID NO. 83)
hsa-miR-4529-5p	MIMAT0019236	3.453	1.07E-32	0.8208	AGGCCAUCAGCAGUCCAA UGAA (SEQ ID NO. 84)
hsa-miR-6861-5p	MIMAT0027623	0.818	1.20E-32	0.8007	ACUGGGUAGGUGGGGCUC CAGG (SEQ ID NO. 85)
hsa-miR-1469	MIMAT0007347	0.758	2.45E-32	0.8228	CUCGGCGCGGGGCGGGGC UCC (SEQ ID NO. 86)
hsa-miR-619-5p	MIMAT0026622	1.750	2.88E-32	0.8413	GCUGGGAUUACAGGCAUG AGCC (SEQ ID NO. 87)
hsa-miR-4448	MIMAT0018967	2.410	3.95E-32	0.8064	GGCUCCUUGGUCUAGGGG UA (SEQ ID NO. 88)
hsa-miR-4658	MIMAT0019725	2.788	4.02E-32	0.8321	GUGAGUGUGGAUCCUGGA GGAU (SEQ ID NO. 89)
hsa-miR-22-3p	MIMAT0000077	2.815	5.70E-32	0.8327	AAGCUGCCAGUUGAAGAA CUGU (SEQ ID NO. 90)
hsa-miR-4776-5p	MIMAT0019932	2.510	6.41E-32	0.8355	GUGGACCAGGAUGGCAAG GGCU (SEQ ID NO. 91)
hsa-miR-320e	MIMAT0015072	3.365	1.05E-31	0.8191	AAAGCUGGGUUGAGAAGG (SEQ ID NO. 92)
hsa-miR-1225-3p	MIMAT0005573	0.741	1.99E-31	0.8297	UGAGCCCCUGUGCCGCCCC CAG (SEQ ID NO. 93)
hsa-miR-6875-5p	MIMAT0027650	-0.840	2.62E-31	0.8291	UGAGGGACCCAGGACAGG AGA (SEQ ID NO. 94)
hsa-miR-4534	MIMAT0019073	1.324	3.60E-31	0.8167	GGAUGGAGGAGGGGUCU (SEQ ID NO. 95)
hsa-miR-4652-5p	MIMAT0019716	3.280	3.60E-31	0.8156	AGGGGACUGGUUAAUAGA ACUA (SEQ ID NO. 96)
hsa-miR-648	MIMAT0003318	3.145	4.13E-31	0.8014	AAGUGUGCAGGGCACUGG U (SEQ ID NO. 97)
hsa-miR-4259	MIMAT0016880	2.262	4.13E-31	0.8147	CAGUUGGGUCUAGGGGUC AGGA (SEQ ID NO. 98)
hsa-miR-107	MIMAT0000104	3.642	6.38E-31	0.8167	AGCAGCAUUGUACAGGGC UAUCA (SEQ ID NO. 99)

hsa-miR-650	MIMAT0003320	2.399	7.82E-31	0.8214	AGGAGGCAGCGCUCUCAG GAC (SEQ ID NO. 100)
-------------	--------------	-------	----------	--------	--

[0151] 3.2. Development of Diagnostic Model

[0152] Diagnostic model development was performed in the discovery set of the lung cancer dataset, which included 208 lung cancer patients and 208 non-cancer controls (FIG. 1A). limma analysis was used to evaluate the statistical significance of differential miRNA expression between lung cancer patients and non-cancer controls. The top 100 differentially expressed miRNAs were listed in Table 1. Ten-fold cross validation showed that a diagnostic model with the top 4 miRNAs ranked by adjusted p values (hsa-miR-5100, hsa-miR-1343-3p, hsa-miR-1290, and hsa-miR-4787-3p) would result in the best AUC in the ROC curve analysis (FIG. 2A). A diagnostic index calculated by the weighted sum of the 4 miRNA expression levels and normalized to the range of zero to ten showed a near-perfect AUC value of 0.999 (FIG. 2B), numerically better than the AUC of 0.993 for the original 2-miRNA model from the original publication (Asakura et al. 2020) ($p = 0.16$). The cut-point of six was chosen to ensure no misclassification of the non-cancer controls in the discovery set to minimize the false positives, which resulted in 98% sensitivity and 100% specificity (FIG. 2C), compared to 99% for both sensitivity and specificity for the original 2-miRNA model (Asakura et al. 2020).

[0153] 3.3. Validation of the Diagnostic Model in the Lung Cancer Validation Set

[0154] The performance of the 4-miRNA model was evaluated in the lung cancer validation set ($n = 3328$), including 1358 lung cancer patients and 1970 non-cancer controls. The 4-miRNA model achieved an AUC of 0.999 (FIG. 2D), significantly better than the AUC of 0.996 for the original 2-miRNA model (Asakura et al. 2020) ($p = 0.01$). The new model also resulted in 99% for both sensitivity and specificity (FIG. 2E), whereas the original 2-miRNA model showed 95% sensitivity and 99% specificity (Asakura et al. 2020).

[0155] Furthermore, the performance of the 4-miRNA model was assessed in clinical subsets of the validation set, as defined by clinical stage, T stage, N stage, M stage, and Histology. Across all clinical subsets, the 4-miRNA model showed sensitivities of approximately 99% or above (FIG. 2G, Table 2), which were superior to the sensitivities of the original 2-miRNA model (Table 2). In particular for early stage lung cancer, e.g., for both patients with stage I lung cancer and patients with T1 tumors, the 4-miRNA model demonstrated >99% sensitivity (FIG. 2G, Table 2), compared to the sensitivities of 95.4 and 95.9%, respectively, for the 2-miRNA model (Table 2). In the prevalent histological types of adenocarcinoma and squamous cell carcinoma, the 4-miRNA model also demonstrated superior performance (FIG. 2G, Table 2), compared to the original 2-miRNA model (Table 2).

Table 2. Comparison of sensitivities in the clinical subsets of the lung cancer validation set between the original 2-miRNA model and the new 4-miRNA model, while maintaining a specificity of >99%.

Clinical Subsets		N	Original 2-miRNA model	New 4-miRNA model	P value*
Clinical Stage	IA	686	96.1%	99.6%	<0.001
	IB	285	93.7%	99.6%	<0.001
	IIA	146	97.3%	97.9%	0.99
	IIB	61	96.7%	98.4%	0.99
	IIIA	164	90.2%	99.4%	<0.001
	IIIB	6	83.3%	100.0%	0.99
	IV	8	100.0%	100.0%	1.00
T Stage	T1a	466	96.1%	99.6%	<0.001
	T1b	297	95.6%	99.3%	0.003
	T2a	435	93.6%	99.1%	<0.001
	T2b	52	92.3%	100.0%	0.134
	T3	89	94.4%	98.9%	0.221
	T4	17	94.1%	100.0%	0.99
N Stage	N0	1047	95.5%	99.5%	<0.001
	N1	166	95.8%	98.2%	0.289
	N2	142	90.1%	99.3%	<0.001
M Stage	M0	1348	94.7%	99.3%	<0.001
	M1a	8	100.0%	100.0%	1.00
Histology	ADC	1038	95.1%	99.2%	<0.001
	SqCC	205	94.2%	99.5%	0.006
	LCC	34	97.1%	100.0%	0.99
	SCLC	22	90.9%	100.0%	0.480
	Others	57	96.5%	100.0%	0.480

* p values calculated by McNemar Test.

5 [0156] Data on paired serum samples (pre- vs. post-surgery) were also available for 180 patients. The diagnostic indices of the 4-miRNA model for post-surgery serum samples were reduced to normal levels below the diagnostic index cut-point (FIG. 2F).

[0157] 3.4. Application of the Diagnostic Model in Additional Cancer Types

10 [0158] The performance of the 4-miRNA model was further assessed in the combined dataset of 3792 patients, including 2038 cancer patients across 12 different cancer types and 1754 non-cancer controls. The bladder, liver, and ovarian cancers had the largest sample sizes with >300 patients in each. Except for breast cancer in which the 4-miRNA model did not perform, the 4-miRNA model showed very strong performances with AUCs >0.95 in biliary tract, bladder, colorectal, esophageal, gastric, glioma, liver, ovarian, pancreatic, and prostate cancers, and an
 15 AUC of 0.876 in Sarcoma (FIG. 3A). Accordingly, the 4-miRNA model demonstrated high sensitivities in the range from 83.2 to 100% for biliary tract, bladder, colorectal, esophageal, gastric, glioma, liver, pancreatic, and prostate cancers, and reasonable sensitivities of 68.2 and 72.0% for ovarian cancer and sarcoma, respectively (FIG. 3B). In addition, for the 1754 non-cancer controls

independent of those included in the lung cancer dataset, the 4-miRNA model maintained a high specificity of 99.3%.

[0159] A further sensitivity analysis with an alternative diagnostic index cut-point of 5.1 that would lower the specificity to 95% resulted in increased sensitivities across all 11 cancer types, demonstrating sensitivities of >90% across ten cancer types with the exception of 76.5% sensitivity for sarcoma (Table 3).

Table 3. Comparison of sensitivities of the 4-miRNA diagnostic model in additional cancer datasets based on the default cut-point vs. alternative cut-point that resulted in 95% specificity.

	Default cut-point based on 99% specificity	Alternative cut-point based on 95% specificity
Biliary Tract Cancer	97.5%	100.0%
Bladder Cancer	98.2%	99.2%
Colorectal Cancer	85.8%	91.6%
Esophageal Cancer	84.7%	95.2%
Gastric Cancer	100.0%	100.0%
Glioma	87.5%	97.5%
Liver Cancer	84.2%	92.5%
Ovarian Cancer	68.2%	90.1%
Pancreatic Cancer	83.2%	95.3%
Prostate Cancer	92.5%	97.5%
Sarcoma	72.0%	76.5%

10

[0160] 4. Discussion

[0161] In this example, we report on the development and performance evaluation of a 4-miRNA diagnostic model for multi-cancer early detection. We demonstrated that in the large independent validation set of 7120 participants including 3396 cancer patients and 3724 non-cancer individuals, the 4-miRNA model can detect 12 cancer types (biliary tract, bladder, colorectal, esophageal, gastric, glioma, live, lung, ovarian, pancreatic, prostate, and sarcoma) simultaneously with high sensitivities (80–100% for ten cancer types, and ~70% for two cancer types) while still maintaining a very high specificity of 99% that is typically required for a screening test to be useful in at-risk general population. To our knowledge, this is the first MCED diagnostic model based on circulating cell-free miRNAs. It is interesting to note that the diagnostic index for lung cancer patients decreased to the levels of non-cancer controls after tumor resection, suggesting that the diagnostic model might have the potential for monitoring tumor recurrence.

[0162] Noninvasive screening tests analyzing circulating nucleic acids and/or proteins have become the driving force of the MCED campaign with significant progress being made recently. Nearly all of the tests that are being developed for MCED are based on the evaluation of circulating

25

tumor DNAs, and most utilize next generation bisulfite sequencing technology to evaluate the methylation patterns of these tumor DNAs (Klein et al. 2021; Cohen et al. 2018; Chen et al. 2020; and Cristiano et al. 2019). Two such tests, Galleri and PanSeer, are developed as methylation-based epigenetic signatures (Klein et al. 2021; Chen 2020). In the analysis of the case-control study of the Circulating Cell-free Genome Atlas (CCGA), Galleri interrogated >100,000 methylated regions and showed that the sensitivity for 12 pre-specified cancers (anus, bladder, colon/rectum, esophagus, head and neck, liver/bile-duct, lung, lymphoma, ovary, pancreas, plasma cell neoplasm, stomach) was 67.6% for patients with stage I–III disease (n = 874) and increased to 76.3% (n = 1346) when stage IV cancer was included, while reaching a 99.3% specificity based on 1254 non-cancer controls (Klein et al. 2021). On the other hand, PanSeer assay which targeted only 477 methylated genomic regions retrospectively analyzed plasma samples from a group of asymptomatic individuals enrolled in a longitudinal cancer monitoring study, and demonstrated a high sensitivity of 95% in 98 individuals who later were diagnosed with one of five cancers (stomach, esophageal, colorectal, lung, and liver cancer) within four years of blood draw (pre-diagnosis samples), but with a lower specificity of 96% in 207 healthy controls (Chen et al. 2020). However, what was puzzling with PanSeer was that when it was evaluated in 113 post-diagnosis plasma samples, the test only showed a lower 88% sensitivity (Chen et al. 2020). Another test called DELFI, based on the genome-wide analysis of cell-free DNA fragmentation patterns by next generation sequencing, achieved a 73% sensitivity across seven cancers (n = 208, breast, bile duct, colorectal, gastric, lung, ovarian, and pancreatic) and 98% specificity (n = 215) (Cristiano et al. 2019). Finally, CancerSEEK, a test combining the measurement of nine protein biomarkers and detection of mutations of 16 genes in circulating cell-free DNA, showed ten-fold cross-validations and median 70% sensitivity (n = 1005) across eight cancers (n = 1005, ovary, liver, stomach, pancreas, esophagus, colorectum, lung, and breast) and 99% specificity (n = 812) (Cohen et al. 2018). In summary, the current MCED tests in development generally showed sensitivities in the range of 60–70% when a high specificity of 99% was mandated. Compared to these tests, our diagnostic model was much simpler with only 4 miRNAs and yet demonstrated substantially higher sensitivities in the range of 80–100% for 10 out of 12 cancer types studied with a large cohort of over 7000 participants. It is worthy of note that a simple diagnostic model not only costs significantly lower, but also can be developed into an in vitro diagnostic (IVD) test using conventional technology platform such as RT-PCR capable of decentralized testing, which has an advantage over NGS-based tests that are usually implemented as a laboratory developed test (LDT). These characteristics are important to drive the wide adoption and compliance of MCED tests as they are intended to target high-risk or at-risk general public.

[0163] Among the 13 cancer types examined in this study, only breast cancer was not detected successfully by the 4-miRNA diagnostic model. While the reason for this underperformance was not clear, it may indicate that breast cancer has a different miRNA expression profile and/or different shedding pattern of miRNAs into the bloodstream. Interestingly, Galleri and CancerSEEK also showed poor sensitivity of 30.5 and 33% in breast cancer, respectively (Klein et al. 2021; Cohen et al. 2018). Nevertheless, the poor performance in breast cancer may not be clinically important because mammography screening has been very effective in detecting early stage breast cancer and decreasing breast cancer mortality (Nelson et al. 2016).

[0164] The ultimate diagnostic performance and clinical value of these MCED tests has to be established in large prospective screening trials with asymptomatic individuals. In the DETECT-A trial enrolling more than 10,000 asymptomatic women, 96 cancers were identified across ten cancer types, CancerSEEK showed a sensitivity of 27%, and that increased to 52% when adding those detected by standard-of-care screening tests (Lennon et al. 2020). In addition, CancerSEEK, when combined by PET-CT scan, showed a specificity of 99.6% and a positive predictive value (PPV) of 40.6%. On the other hand, in the interim analysis of 4033 participants from the prospective PATHFINDER study of Galleri test, 40 had a positive test result, 18 of them were confirmed to have cancer leading to a PPV of 45% (Beer et al. 2021). For our 4-miRNA diagnostic model, assuming 1% cancer incidence rate and a conservative average sensitivity of 85 and 99.3% specificity, our model would provide a PPV of 55% when screening asymptomatic individuals. This is significantly higher than the PPVs for the four USPSTF recommended single cancer screenings, which range from 3.7 to 4.4% (Lehman et al. 2017; U.S. Food and Drug Administration Cologuard Summary of Safety and Effectiveness Data, 2014; and National Lung Screening Trial Research Team, 2013).

[0165] 5. Conclusions

[0166] In summary, our study has provided proof-of-concept data for a simple and affordable blood-based diagnostic test that detects multiple cancers. The 12 cancer types that were detected in this study account for almost 380,000 (~62%) estimated cancer deaths in the US in 2021. While the early detection of these cancers should conceivably reduce the cancer-related deaths, the ultimate determination of clinical performance and clinical utility will require the evaluation in large prospective studies with asymptomatic individuals from the intended use population.

[0167] It is noted that the although the examples and data provided above only cover 12 cancers, for which the miRNA biomarker set, especially the 4-miRNA biomarker set, has demonstrated excellent power in the detection of cancers with very high accuracy, there is no limitation to the cancer types that the miRNA biomarker set can be applied. Accordingly, the scope

of the present disclosure shall be interpreted to cover other cancer types as well. The fact that the model provided in the present disclosure works for 12 of the 13 cancer types studied strongly suggests that the method is applicable to most, if not all of the cancer types.

5 REFERENCES

- Ritchie, ME; et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* 43(7), e47.
- 10 Venables, WN and Ripley, BD (2002) *Modern Applied Statistics with S*. Fourth edition. Springer.
- Tibshirani, R (1996). "Regression Shrinkage and Selection via the lasso". *Journal of the Royal Statistical Society. Series B (methodological)*. Wiley. **58** (1): 267–88.
- 15 Hoerl, AE and Kennard, RW (1970). "Ridge Regression: Biased Estimation for Nonorthogonal problems". *Technometrics*. **12** (1): 55–67.
- Ripley, BD (1996) *Pattern Recognition and Neural Networks*. Cambridge University Press.
- 20 Kozomara, A and Griffiths-Jones, S (2010). "MiRBase: integrating microRNA annotation and deep-sequencing data". *Nucleic Acids Research*. 39 (Database issue): D152–7.
- miRBase: the microRNA database: <http://www.mirbase.org/>
- 25 The Bioconductor Open Source Software For Bioinformatics: <http://www.bioconductor.org>
- The R Project for Statistical Computing: <https://www.r-project.org/>
- Asakura, K; et al. (2020). A MiRNA-Based Diagnostic Model Predicts Resectable Lung Cancer in Humans with High Accuracy. *Commun. Biol.* 3, 134.
- 30 Yokoi, A; et al. (2018). Integrated Extracellular MicroRNA Profiling for Ovarian Cancer Screening. *Nat. Commun.* 9, 4319.
- 35 Usuba, W; et al. (2019). Circulating MiRNA Panels for Specific and Early Detection in Bladder Cancer. *Cancer Sci.* 110, 408–419.
- Yamamoto, Y; et al. (2020). Highly Sensitive Circulating MicroRNA Panel for Accurate Detection of Hepatocellular Carcinoma in Patients With Liver Disease. *Hepatol. Commun.* 4, 284–297.
- 40 Klein, EA; et al. (2021). Clinical Validation of a Targeted Methylation-Based Multi-Cancer Early Detection Test Using an Independent Validation Set. *Ann. Oncol.: Off. J. Eur. Soc. Med. Oncol.* 32, 1167–1177.
- 45 Cohen, JD; et al. (2018). Detection and Localization of Surgically Resectable Cancers with a Multi-Analyte Blood Test. *Science*. 359, 926–930.
- Chen, X; et al. (2020). Non-Invasive Early Detection of Cancer Four Years before Conventional Diagnosis Using a Blood Test. *Nat. Commun.* 11, 3475.
- 50

- Cristiano, S; et al. (2019). Genome-Wide Cell-Free DNA Fragmentation in Patients with Cancer. *Nature*. 570, 385–389.
- 5 Nelson, HD; et al. (2016). Effectiveness of Breast Cancer Screening: Systematic Review and Meta-Analysis to Update the 2009 U.S. Preventive Services Task Force Recommendation. *Ann. Intern. Med.* 164, 244–255.
- 10 Lennon, AM; et al. (2020). Feasibility of Blood Testing Combined with PET-CT to Screen for Cancer and Guide Intervention. *Science*. 369, eabb9601.
- Beer, T; et al. (2021). Interim Results of PATHFINDER, a Clinical Use Study Using a Methylation-Based Multi-Cancer Early Detection Test. *J. Clin. Oncol.* 39, 3010.
- 15 Lehman, CD; et al. (2017). National Performance Benchmarks for Modern Screening Digital Mammography: Update from the Breast Cancer Surveillance Consortium. *Radiology*. 283, 49–58.
- U.S. Food and Drug Administration Cologuard Summary of Safety and Effectiveness Data (Premarket Approval Application P130017); 2014.
- 20 National Lung Screening Trial Research Team; Church, TR; et al. (2013). Results of Initial Low-Dose Computed Tomographic Screening for Lung Cancer. *New Engl. J. Med.* 2013, 368, 1980–1991.
- Nielsen, PE; et al. (1991). Sequence-selective recognition of DNA by strand displacement with a thymine-substituted polyamide. *Science*. 254, p. 1497-500.
- 25 Obika, S; et al. (1998). Stability and structural features of the duplexes containing nucleoside analogues with a fixed N-type conformation, 2'-O,4'- C-methyleneribonucleosides. *Tetrahedron Lett.* 39, p. 5401-5404.
- 30 Green, MR and Sambrook, J. (2012). *Molecular Cloning: A Laboratory Manual*, 4th Ed., Cold Spring Harbor Press, Cold Spring Harbor, N.Y.
- Sambrook, J; et al. (1989). *Molecular Cloning: A Laboratory Manual*, New York: Cold Spring Harbor Press.
- 35 Zhang, Z; et al. (2000). A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* 7, p. 203-214.
- Altschul, SF; et al. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, Vol. 40 215, p. 403-410.
- Pearson, WR et al. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U. S. A.*, Vol. 85, p. 2444-2448.
- 45 Yun, SJ; et al. (2012). Cell-free microRNAs in urine as diagnostic and prognostic biomarkers of bladder cancer. *Int J Oncol.* 2012 Nov;41(5):1871-8.
- Park, NJ; et al. (2009). Salivary microRNA: discovery, characterization, and clinical utility for oral cancer detection. *Clin Cancer Res.* 2009 Sep 1;15(17):5473-7.
- 50

CLAIMS

1. A method for detecting a cancer from a biological sample obtained from a subject, comprising:
determining an expression profile of an miRNA biomarker set consisting of at least one
5 miRNA from the biological sample, wherein the miRNA biomarker set comprises hsa-miR-5100;
calculating a diagnostic index of the biological sample based on the expression profile of the
miRNA biomarker set, wherein the diagnostic index is calculated based on formula:

$$\text{diagnostic index} = \sum_{i=1}^n t_i * \text{miRNA}_i; \quad (I)$$

10 where n is the total number of the at least one miRNA in the miRNA biomarker set, miRNA_i is
the expression level of i^{th} miRNA in the miRNA biomarker set, i is an integer greater than zero
and smaller than or equal to n ; and t_i is a weight for the i^{th} miRNA; and
classifying the subject as having the cancer or not based on the calculated diagnostic index,
wherein the subject is classified as having the cancer if the calculated diagnostic index is greater
15 than or equal to a pre-determined threshold or as not having the cancer if otherwise;
wherein:
the method is capable of achieving diagnostic accuracy having an AUC value greater than
approximately 0.780.
- 20 2. The method of claim 1, wherein the miRNA biomarker set further comprises one or more of
hsa-miR-1343-3p, hsa-miR-1290, hsa-miR-4787-3p, hsa-miR-6877-5p, hsa-miR-17-3p, hsa-miR-
6765-5p, hsa-miR-1268b, hsa-miR-4258, hsa-miR-451a, hsa-miR-1228-5p, hsa-miR-8073, hsa-
miR-4454, hsa-miR-187-5p, hsa-miR-4286, hsa-miR-6746-5p, hsa-miR-663b, hsa-miR-6075,
hsa-miR-5001-5p, hsa-miR-6789-5p, hsa-miR-4513, hsa-miR-3192-5p, hsa-miR-8060, hsa-miR-
25 668-5p, hsa-miR-1268a, hsa-miR-1273g-3p, hsa-miR-4706, hsa-miR-124-3p, hsa-miR-1260b,
hsa-miR-4740-5p, hsa-miR-320b, hsa-miR-7977, hsa-miR-29b-3p, hsa-miR-4708-3p, hsa-miR-
4525, hsa-miR-92b-3p, hsa-miR-4257, hsa-miR-4727-3p, hsa-miR-92a-3p, hsa-miR-663a, hsa-
miR-6787-5p, hsa-miR-3131, hsa-miR-6802-5p, hsa-miR-654-5p, hsa-miR-6511b-5p, hsa-miR-
29b-1-5p, hsa-miR-4417, hsa-miR-4736, hsa-miR-6840-3p, hsa-miR-4710, hsa-miR-4635, hsa-
30 miR-296-3p, hsa-miR-1199-5p, hsa-miR-7975, hsa-miR-4480, hsa-miR-3648, hsa-miR-371a-5p,
hsa-miR-4771, hsa-miR-6717-5p, hsa-miR-1254, hsa-miR-1246, hsa-miR-23b-3p, hsa-miR-320a,
hsa-miR-4687-5p, hsa-miR-191-5p, hsa-miR-320c, hsa-miR-6131, hsa-miR-4515, hsa-miR-342-
5p, hsa-miR-4718, hsa-miR-23a-3p, hsa-miR-4455, hsa-miR-211-3p, hsa-miR-3122, hsa-miR-
103a-3p, hsa-miR-4429, hsa-miR-920, hsa-miR-3194-3p, hsa-miR-4754, hsa-miR-1238-5p, hsa-
35 miR-3191-3p, hsa-miR-4755-3p, hsa-miR-3688-5p, hsa-miR-4529-5p, hsa-miR-6861-5p, hsa-
miR-1469, hsa-miR-619-5p, hsa-miR-4448, hsa-miR-4658, hsa-miR-22-3p, hsa-miR-4776-5p,

hsa-miR-320e, hsa-miR-1225-3p, hsa-miR-6875-5p, hsa-miR-4534, hsa-miR-4652-5p, hsa-miR-648, hsa-miR-4259, hsa-miR-107, and hsa-miR-650.

3. The method of claim 1, wherein the miRNA biomarker set further comprises one or more of
5 hsa-miR-1343-3p, hsa-miR-1290, hsa-miR-4787-3p, hsa-miR-6877-5p, hsa-miR-17-3p, hsa-miR-6765-5p, hsa-miR-1268b, hsa-miR-4258, hsa-miR-451a, hsa-miR-1228-5p, hsa-miR-8073, hsa-miR-4454, hsa-miR-187-5p, hsa-miR-4286, hsa-miR-6746-5p, hsa-miR-663b, hsa-miR-6075, hsa-miR-5001-5p, hsa-miR-6789-5p, hsa-miR-4513, hsa-miR-3192-5p, hsa-miR-8060, hsa-miR-668-5p, hsa-miR-1268a, hsa-miR-1273g-3p, hsa-miR-4706, hsa-miR-124-3p, hsa-miR-1260b,
10 hsa-miR-4740-5p, hsa-miR-320b, hsa-miR-7977, hsa-miR-29b-3p, hsa-miR-4708-3p, hsa-miR-4525, hsa-miR-92b-3p, hsa-miR-4257, hsa-miR-4727-3p, hsa-miR-92a-3p, hsa-miR-663a, hsa-miR-6787-5p, hsa-miR-3131, hsa-miR-6802-5p, hsa-miR-654-5p, hsa-miR-6511b-5p, hsa-miR-29b-1-5p, hsa-miR-4417, hsa-miR-4736, hsa-miR-6840-3p, and hsa-miR-4710.

15 4. The method of claim 1, wherein the miRNA biomarker set further comprises one or more of hsa-miR-1343-3p, hsa-miR-1290, hsa-miR-4787-3p, hsa-miR-6877-5p, hsa-miR-17-3p, hsa-miR-6765-5p, hsa-miR-1268b, hsa-miR-4258, hsa-miR-451a, hsa-miR-1228-5p, hsa-miR-8073, hsa-miR-4454, hsa-miR-187-5p, hsa-miR-4286, hsa-miR-6746-5p, hsa-miR-663b, hsa-miR-6075, hsa-miR-5001-5p, and hsa-miR-6789-5p.

20 5. The method of claim 4, wherein the miRNA biomarker set consists of hsa-miR-5100, hsa-miR-1343-3p, hsa-miR-1290, hsa-miR-4787-3p, hsa-miR-6877-5p, hsa-miR-17-3p, hsa-miR-6765-5p, hsa-miR-1268b, hsa-miR-4258, hsa-miR-451a, hsa-miR-1228-5p, hsa-miR-8073, hsa-miR-4454, hsa-miR-187-5p, hsa-miR-4286, hsa-miR-6746-5p, hsa-miR-663b, hsa-miR-6075, hsa-miR-5001-5p, and hsa-miR-6789-5p.
25

6. The method of claim 1, wherein the miRNA biomarker set further comprises one or more of hsa-miR-1343-3p, hsa-miR-1290, and hsa-miR-4787-3p.

30 7. The method of claim 6, wherein the miRNA biomarker set consists of hsa-miR-5100, hsa-miR-1343-3p, hsa-miR-1290, and hsa-miR-4787-3p.

8. The method of claim 7, wherein the method is capable of achieving diagnostic accuracy having an AUC value greater than approximately 0.850.

35

9. The method of claim 8, wherein the cancer is selected from a group consisting of lung cancer, biliary tract cancer, bladder cancer, colorectal cancer, esophageal cancer, gastric cancer, glioma cancer, liver cancer, pancreatic cancer, prostate cancer, ovarian cancer, and sarcoma.
- 5 10. The method of claim 8, wherein the method is capable of achieving diagnostic accuracy having an AUC value greater than approximately 0.950.
11. The method of claim 10, wherein the cancer is selected from a group consisting of lung cancer, biliary tract cancer, bladder cancer, colorectal cancer, esophageal cancer, gastric cancer, glioma
10 cancer, liver cancer, ovarian cancer, pancreatic cancer, and prostate cancer.
12. The method of claim 10, wherein the method is capable of achieving diagnostic accuracy having an AUC value greater than approximately 0.990.
- 15 13. The method of claim 12, wherein the cancer is selected from a group consisting of lung cancer, biliary tract cancer, bladder cancer, esophageal cancer, gastric cancer, glioma cancer, and prostate cancer.
14. The method of claim 12, wherein the method is capable of achieving a diagnostic accuracy
20 having an AUC value greater than approximately 0.999.
15. The method of claim 14, wherein the cancer is selected from a group consisting of lung cancer, and gastric cancer.
- 25 16. The method of claim 7, wherein the method is capable of achieving diagnostic accuracy having a sensitivity greater than approximately 68.0% while having a specificity greater than approximately 99.0%.
17. The method of claim 16, wherein the cancer is selected from a group consisting of lung cancer,
30 biliary tract cancer, bladder cancer, colorectal cancer, esophageal cancer, gastric cancer, glioma cancer, liver cancer, pancreatic cancer, prostate cancer, ovarian cancer, and sarcoma.
18. The method of claim 16, wherein the method is capable of achieving diagnostic accuracy having a sensitivity greater than approximately 83.0% while having a specificity greater than
35 approximately 99.0%.

19. The method of claim 18, wherein the cancer is selected from a group consisting of lung cancer, biliary tract cancer, bladder cancer, colorectal cancer, esophageal cancer, gastric cancer, glioma cancer, liver cancer, pancreatic cancer, and prostate cancer.
- 5 20. The method of claim 18, wherein the method is capable of achieving diagnostic accuracy having a sensitivity greater than approximately 99.0% and having a specificity greater than approximately 99.0%.
21. The method of claim 20, wherein the cancer is selected from a group consisting of lung cancer
10 and gastric cancer.
22. The method of any of claims 1-21, wherein in the calculating the diagnostic index of the biological sample based on the expression profile of the miRNA biomarker set, the diagnostic index is calculated via an unweighted model.
- 15 23. The method of any of claims 1-21, wherein in the calculating the diagnostic index of the biological sample based on the expression profile of the miRNA biomarker set, the diagnostic index is calculated via a weighted model using weights from one selected from a group consisting of Linear Models for Microarray Data (limma) model, logistic regression model, linear
20 discriminant analysis (LDA) model, conditional logistic regression model, lasso regression model, ridge regression model, random forest, support vector machine, and probit regression model.
24. The method of claim 23, wherein the diagnostic index is calculated via a weighted model using weights from the limma model.
- 25 25. The method of any one of claims 1-24, wherein the pre-determined threshold is 1110, and the method is capable of achieving diagnostic accuracy having a specificity value greater than approximately 0.95.
- 30 26. The method of any one of claims 1-24, wherein the pre-determined threshold is 1200, and the method is capable of achieving diagnostic accuracy having a specificity value greater than approximately 0.99.
27. The method of any of claims 1-26, further comprising, after the calculating a diagnostic index
35 of the biological sample and before the classifying the subject as having the cancer or not:
obtaining a normalized diagnostic index based on the calculated diagnostic index;
wherein:

the classifying the subject as having the cancer or not based on the calculated diagnostic index comprises:

classifying the subject as having the cancer if the normalized diagnostic index is equal to or greater than a preset cut-point; or

5 classifying the subject as not having the cancer if otherwise.

28. The method of claim 27, wherein in the obtaining a normalized diagnostic index based on the calculated diagnostic index, the normalized diagnostic index is calculated based on formula:

$$\text{normalized diagnostic index} = \frac{\text{diagnostic index} - \text{param}_{\text{location}}}{\text{param}_{\text{scale}}}, \quad (\text{II})$$

10

where the $\text{param}_{\text{location}}$ and $\text{param}_{\text{scale}}$ are respectively a location parameter and a scale parameter configured to allow the normalized diagnostic index to be within a range no less than a first preset value and no greater than a second preset value.

15 29. The method of claim 28, wherein the diagnostic index is calculated via a weighted model using weights from the limma model, and the first preset value is 0, and the second preset value is 10.

30. The method of claim 29, wherein the preset cut-point is 5.1, and the method is capable of achieving diagnostic accuracy having a specificity value greater than approximately 0.95.

20

31. The method of claim 29, wherein the preset cut-point is 6.0, and the method is capable of achieving diagnostic accuracy having a specificity value greater than approximately 0.99.

25 32. The method of any of claims 1-31, wherein the biological sample is a liquid biopsy sample selected from a group consisting of a blood sample, a serum sample, a plasma sample, a urine sample, a saliva sample, and a spatium sample.

30 33. The method of any of claims 1-32, wherein in the determining an expression profile of an miRNA biomarker set consisting of at least one miRNA from the biological sample, the expression profile of the miRNA biomarker set is obtained by means of at least one of the methods of Northern Blotting, microarray analysis, RNA-sequencing, or RNA in-situ hybridization.

35 34. The method of any of claims 1-32, wherein in the determining an expression profile of an miRNA biomarker set consisting of at least one miRNA from the biological sample, the expression profile of the miRNA biomarker set is obtained by means of a nucleic acid amplification procedure, comprising at least one of reverse-transcription PCR (RT-PCR), quantitative RT-PCR (qRT-PCR), or digital RT-PCR.

35. The method of any one of claims 1-34, further comprising: performing an evaluation of the subject, wherein said evaluation comprises a diagnosis of the cancer or a detection of a recurrence of the cancer.

5

36. The method of any one of claims 1-35, further comprising: administering to the subject a therapeutic regimen when the subject is classified as having the cancer.

37. A kit for detecting a cancer from a biological sample obtained from a subject, comprising at least one nucleic acid and at least one instruction, wherein:

10

each of the at least one nucleic acid is capable of specifically recognizing each miRNA in an miRNA biomarker set to thereby allow an expression profile of the miRNA biomarker set to be obtained from the biological sample, wherein the miRNA biomarker set comprises hsa-miR-5100;

15

the at least one instruction comprises:

a first instruction, comprising a first sub-instruction for calculating a diagnostic index of the biological sample based on the expression profile of the miRNA biomarker set, wherein the diagnostic index is calculated based on formula:

$$\text{diagnostic index} = \sum_{i=1}^n t_i * \text{miRNA}_i; \quad (I)$$

20

where n is the total number of the at least one miRNA in the miRNA biomarker set, miRNA_i is the expression level of i^{th} miRNA in the miRNA biomarker set, i is an integer greater than zero and smaller than or equal to n , and t_i is a weight for the i^{th} miRNA; and

25

a second instruction for classifying the subject as having the cancer or not, wherein the subject is classified as having the cancer if the calculated diagnostic index is greater than or equal to a pre-determined threshold or as not having the cancer if otherwise.

38. The kit of claim 37, wherein the at least one nucleic acid comprises a polynucleotide capable of specifically hybridizing under a stringent condition to:

30

(a) a polynucleotide comprising or consisting of a nucleotide sequence of SEQ ID NO: 1, a derivative thereof, a variant thereof having at least 80% sequence identity, or a fragment thereof comprising 15 or more consecutive nucleotides; or

(b) a polynucleotide comprising or consisting of a nucleotide sequence complementary to a nucleotide sequence of SEQ ID NO: 1, a derivative thereof, a variant thereof having at least 80% sequence identity, or a fragment thereof comprising 15 or more consecutive nucleotides.

35

39. The kit of claim 37 or claim 38, wherein the miRNA biomarker set further comprises one or more of hsa-miR-1343-3p, hsa-miR-1290, hsa-miR-4787-3p, hsa-miR-6877-5p, hsa-miR-17-3p, hsa-miR-6765-5p, hsa-miR-1268b, hsa-miR-4258, hsa-miR-451a, hsa-miR-1228-5p, hsa-miR-8073, hsa-miR-4454, hsa-miR-187-5p, hsa-miR-4286, hsa-miR-6746-5p, hsa-miR-663b, hsa-miR-6075, hsa-miR-5001-5p, hsa-miR-6789-5p, hsa-miR-4513, hsa-miR-3192-5p, hsa-miR-8060, hsa-miR-668-5p, hsa-miR-1268a, hsa-miR-1273g-3p, hsa-miR-4706, hsa-miR-124-3p, hsa-miR-1260b, hsa-miR-4740-5p, hsa-miR-320b, hsa-miR-7977, hsa-miR-29b-3p, hsa-miR-4708-3p, hsa-miR-4525, hsa-miR-92b-3p, hsa-miR-4257, hsa-miR-4727-3p, hsa-miR-92a-3p, hsa-miR-663a, hsa-miR-6787-5p, hsa-miR-3131, hsa-miR-6802-5p, hsa-miR-654-5p, hsa-miR-6511b-5p, hsa-miR-29b-1-5p, hsa-miR-4417, hsa-miR-4736, hsa-miR-6840-3p, hsa-miR-4710, hsa-miR-4635, hsa-miR-296-3p, hsa-miR-1199-5p, hsa-miR-7975, hsa-miR-4480, hsa-miR-3648, hsa-miR-371a-5p, hsa-miR-4771, hsa-miR-6717-5p, hsa-miR-1254, hsa-miR-1246, hsa-miR-23b-3p, hsa-miR-320a, hsa-miR-4687-5p, hsa-miR-191-5p, hsa-miR-320c, hsa-miR-6131, hsa-miR-4515, hsa-miR-342-5p, hsa-miR-4718, hsa-miR-23a-3p, hsa-miR-4455, hsa-miR-211-3p, hsa-miR-3122, hsa-miR-103a-3p, hsa-miR-4429, hsa-miR-920, hsa-miR-3194-3p, hsa-miR-4754, hsa-miR-1238-5p, hsa-miR-3191-3p, hsa-miR-4755-3p, hsa-miR-3688-5p, hsa-miR-4529-5p, hsa-miR-6861-5p, hsa-miR-1469, hsa-miR-619-5p, hsa-miR-4448, hsa-miR-4658, hsa-miR-22-3p, hsa-miR-4776-5p, hsa-miR-320e, hsa-miR-1225-3p, hsa-miR-6875-5p, hsa-miR-4534, hsa-miR-4652-5p, hsa-miR-648, hsa-miR-4259, hsa-miR-107, and hsa-miR-650.

20

40. The kit of claim 39, wherein the at least one nucleic acid further comprises at least one polynucleotide, each capable of specifically hybridizing under a stringent condition to:

(a) a polynucleotide comprising or consisting of a nucleotide sequence of any one of SEQ ID NOS: 2-100, a derivative thereof, a variant thereof having at least 80% sequence identity, or a fragment thereof comprising 15 or more consecutive nucleotides; or

(b) a polynucleotide comprising or consisting of a nucleotide sequence complementary to a nucleotide sequence of any one of SEQ ID NOS: 2-100, a derivative thereof, a variant thereof having at least 80% sequence identity, or a fragment thereof comprising 15 or more consecutive nucleotides.

30

41. The kit of claim 37 or claim 38, wherein the miRNA biomarker set further comprises one or more of hsa-miR-1343-3p, hsa-miR-1290, hsa-miR-4787-3p, hsa-miR-6877-5p, hsa-miR-17-3p, hsa-miR-6765-5p, hsa-miR-1268b, hsa-miR-4258, hsa-miR-451a, hsa-miR-1228-5p, hsa-miR-8073, hsa-miR-4454, hsa-miR-187-5p, hsa-miR-4286, hsa-miR-6746-5p, hsa-miR-663b, hsa-miR-6075, hsa-miR-5001-5p, hsa-miR-6789-5p, hsa-miR-4513, hsa-miR-3192-5p, hsa-miR-8060,

35

hsa-miR-668-5p, hsa-miR-1268a, hsa-miR-1273g-3p, hsa-miR-4706, hsa-miR-124-3p, hsa-miR-1260b, hsa-miR-4740-5p, hsa-miR-320b, hsa-miR-7977, hsa-miR-29b-3p, hsa-miR-4708-3p, hsa-miR-4525, hsa-miR-92b-3p, hsa-miR-4257, hsa-miR-4727-3p, hsa-miR-92a-3p, hsa-miR-663a, hsa-miR-6787-5p, hsa-miR-3131, hsa-miR-6802-5p, hsa-miR-654-5p, hsa-miR-6511b-5p, hsa-miR-29b-1-5p, hsa-miR-4417, hsa-miR-4736, hsa-miR-6840-3p, and hsa-miR-4710.

42. The kit of claim 41, wherein the at least one nucleic acid further comprises at least one polynucleotide, each capable of specifically hybridizing under a stringent condition to:

- (a) a polynucleotide comprising or consisting of a nucleotide sequence of any one of SEQ ID NOS: 2-50, a derivative thereof, a variant thereof having at least 80% sequence identity, or a fragment thereof comprising 15 or more consecutive nucleotides; or
- (b) a polynucleotide comprising or consisting of a nucleotide sequence complementary to a nucleotide sequence of any one of SEQ ID NOS: 2-50, a derivative thereof, a variant thereof having at least 80% sequence identity, or a fragment thereof comprising 15 or more consecutive nucleotides.

43. The kit of claim 37 or claim 38, wherein the miRNA biomarker set further comprises one or more of hsa-miR-1343-3p, hsa-miR-1290, hsa-miR-4787-3p, hsa-miR-6877-5p, hsa-miR-17-3p, hsa-miR-6765-5p, hsa-miR-1268b, hsa-miR-4258, hsa-miR-451a, hsa-miR-1228-5p, hsa-miR-8073, hsa-miR-4454, hsa-miR-187-5p, hsa-miR-4286, hsa-miR-6746-5p, hsa-miR-663b, hsa-miR-6075, hsa-miR-5001-5p, and hsa-miR-6789-5p.

44. The kit of claim 43, wherein the at least one nucleic acid further comprises at least one polynucleotide, each capable of specifically hybridizing under a stringent condition to:

- (a) a polynucleotide comprising or consisting of a nucleotide sequence of any one of SEQ ID NOS: 2-20, a derivative thereof, a variant thereof having at least 80% sequence identity, or a fragment thereof comprising 15 or more consecutive nucleotides; or
- (b) a polynucleotide comprising or consisting of a nucleotide sequence complementary to a nucleotide sequence of any one of SEQ ID NOS: 2-20, a derivative thereof, a variant thereof having at least 80% sequence identity, or a fragment thereof comprising 15 or more consecutive nucleotides.

45. The kit of claim 43, wherein the miRNA biomarker set consists of hsa-miR-5100, hsa-miR-1343-3p, hsa-miR-1290, hsa-miR-4787-3p, hsa-miR-6877-5p, hsa-miR-17-3p, hsa-miR-6765-5p, hsa-miR-1268b, hsa-miR-4258, hsa-miR-451a, hsa-miR-1228-5p, hsa-miR-8073, hsa-miR-4454,

hsa-miR-187-5p, hsa-miR-4286, hsa-miR-6746-5p, hsa-miR-663b, hsa-miR-6075, hsa-miR-5001-5p, and hsa-miR-6789-5p.

46. The kit of claim 45, wherein the at least one nucleic acid consists of a total of 20 polynucleotides which are respectively capable of specifically hybridizing under a stringent condition to:

- (a) polynucleotides respectively comprising or consisting of nucleotide sequences of SEQ ID NOS: 1-20, derivatives thereof, variants thereof each having at least 80% sequence identity, or fragments thereof each comprising 15 or more consecutive nucleotides; or
- (b) polynucleotides respectively comprising or consisting of nucleotide sequences which are respectively complementary to nucleotide sequences of SEQ ID NOS: 1-20, derivatives thereof, variants thereof each having at least 80% sequence identity, or fragments thereof each comprising 15 or more consecutive nucleotides.

47. The kit of claim 37 or claim 38, wherein the miRNA biomarker set further comprises one or more of hsa-miR-1343-3p, hsa-miR-1290, and hsa-miR-4787-3p.

48. The kit of claim 47, wherein the at least one nucleic acid further comprises at least one polynucleotide, each capable of specifically hybridizing under a stringent condition to:

- (a) a polynucleotide comprising or consisting of a nucleotide sequence of any one of SEQ ID NOS: 2-4, a derivative thereof, a variant thereof having at least 80% sequence identity, or a fragment thereof comprising 15 or more consecutive nucleotides; or
- (b) a polynucleotide comprising or consisting of a nucleotide sequence complementary to a nucleotide sequence of any one of SEQ ID NOS: 2-4, a derivative thereof, a variant thereof having at least 80% sequence identity, or a fragment thereof comprising 15 or more consecutive nucleotides.

49. The kit of claim 47, wherein the miRNA biomarker set consists of hsa-miR-5100, hsa-miR-1343-3p, hsa-miR-1290, and hsa-miR-4787-3p.

50. The kit of claim 49, wherein the at least one nucleic acid consists of a total of 4 polynucleotides which are respectively capable of specifically hybridizing under a stringent condition to:

- (a) polynucleotides respectively comprising or consisting of nucleotide sequences of SEQ ID NOS: 1-4, derivatives thereof, variants thereof each having at least 80% sequence identity, or fragments thereof each comprising 15 or more consecutive nucleotides; or

(b) polynucleotides respectively comprising or consisting of nucleotide sequences which are respectively complementary to nucleotide sequences of SEQ ID NOS: 1-4, derivatives thereof, variants thereof each having at least 80% sequence identity, or fragments thereof each comprising 15 or more consecutive nucleotides.

5

51. The kit of any one of claims 37-50, wherein in the first sub-instruction of the first instruction, the diagnostic index is calculated via an unweighted model.

10

52. The kit of any one of claims 37-50, wherein in the first sub-instruction of the first instruction, the diagnostic index is calculated via a weighted model using weights from one selected from a group consisting of Linear Models for Microarray Data (limma) model, logistic regression model, linear discriminant analysis (LDA) model, conditional logistic regression model, lasso regression model, ridge regression model, random forest, support vector machine, and probit regression model.

15

53. The kit of claim 52, wherein the diagnostic index is calculated via a weighted model using weights from the limma model.

20

54. The kit of any one of claims 37-53, wherein the pre-determined threshold is 1110, and the second instruction further comprises an indication that classification has a specificity value greater than approximately 0.95.

25

55. The kit of any one of claims 37-53, wherein the pre-determined threshold is 1200, and the second instruction further comprises an indication that classification has a specificity value greater than approximately 0.99.

30

56. The kit of any of claims 37-55, wherein the first instruction further comprises a second sub-instruction for obtaining a normalized diagnostic index based on the diagnostic index calculated according to the first sub-instruction, wherein in the second instruction, the subject is classified as having the cancer if the normalized diagnostic index is greater than or equal to is equal to or greater than a preset cut-point or as not having the cancer if otherwise.

57. The kit of claim 56, wherein in the second sub-instruction, the normalized diagnostic index is calculated based on formula:

35

$$\text{normalized diagnostic index} = \frac{\text{diagnostic index} - \text{param}_{\text{location}}}{\text{param}_{\text{scale}}}, \quad (\text{II})$$

where the $\text{param}_{\text{location}}$ and $\text{param}_{\text{scale}}$ are respectively a location parameter and a scale parameter configured to allow the normalized diagnostic index to be within a range no less than a first preset value and no greater than a second preset value.

5 58. The kit of claim 57, wherein in the first instruction, the diagnostic index is calculated via a weighted model using weights from the limma model, and the first preset value is 0, and the second preset value is 10.

59. The kit of claim 58, wherein the preset cut-point is 5.1, and the second instruction further
10 comprises an indication that classification has a specificity value greater than approximately 0.95.

60. The kit of claim 58, wherein the preset cut-point is 6.0, and the second instruction further comprises an indication that classification has a specificity value greater than approximately 0.95.

15 61. The kit of any one of claims 37-60, wherein the at least one instruction further comprises a third instruction for performing an evaluation of the subject, wherein said evaluation comprises a diagnosis of the cancer or a detection of a recurrence of the cancer.

62. The kit of any one of claims 37-61, wherein the at least one instruction further comprises a
20 fourth instruction for administering to the subject a therapeutic regimen when the subject is classified as having the cancer.

63. The kit of any of claims 37-62, wherein the at least one instruction further comprises a first additional instruction for obtaining the expression profile of the miRNA biomarker set, comprising
25 a procedure for performing Northern Blotting, microarray analysis, RNA-sequencing, or RNA in-situ hybridization by means of the at least one nucleic acid.

64. The kit of claim 63, wherein the at least one nucleic acid is arranged on a molecular array.

30 65. The kit of any one of claims 37-62, further comprising at least one set of amplification primers, each set capable of specifically amplifying each of the at least one miRNA in the miRNA biomarker set from the biological sample.

66. The kit of claim 65, wherein the at least one instruction further comprises a second additional
35 instruction for obtaining the expression profile of the miRNA biomarker set, comprising a procedure for performing reverse-transcription PCR (RT-PCR), quantitative RT-PCR (qRT-PCR),

or digital RT-PCR by means of the at least one nucleic acid and the at least one set of amplification primers.

67. The kit of any of claims 37-66, wherein the biological sample is a liquid biopsy sample selected
5 from a group consisting of a blood sample, a serum sample, a plasma sample, a urine sample, a saliva sample, and a spatium sample.

68. A system for detecting a cancer in a subject, comprising:

a processor; and
10 a non-transitory storage medium containing program instructions for execution by said processor, said program instructions causing said processor to execute steps in the method according to any one of claims 1-36.

69. A non-transitory storage medium, storing computer-executable program instructions which,
15 when executed by a processor, cause the processor to execute the method according to any one of claims 1-36.