



(51) International Patent Classification:
C12Q 1/68 (2006.01)

(21) International Application Number:
PCT/EP2013/062993

(22) International Filing Date:
21 June 2013 (21.06.2013)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/662,769 21 June 2012 (21.06.2012) US

(71) Applicant (for all designated States except US): **PHILIP MORRIS PRODUCTS S.A.** [CH/CH]; Quai Jeanrenaud 3, CH-2000 Neuchâtel (CH).

(72) Inventors; and

(71) Applicants : **BOUE, Stéphanie** [FR/CH]; Verger-Clottu 3, CH-2068 Hauterive (CH). **MARTIN, Florian** [CH/CH]; Chemin de l'Orée 1, CH-Peseux 2034 (CH). **TALIKKA, Marja** [FI/CH]; Rue de la Tournelle 6b, CH-1350 Orbe (CH). **XIANG, Yang** [SG/CH]; Rue du Rocher 24, CH-2000 Neuchâtel (CH).

(74) Agent: **MEIER, Jürgen**; Vossius & Partner, Siebertstrasse 4, 81675 München (DE).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— with international search report (Art. 21(3))

(54) Title: GENE SIGNATURES FOR CLASSIFYING AND GRADING LUNG CANCER

(57) Abstract: The present invention relates to biomarkers and gene signatures that are useful for diagnosing, classifying and prognosing lung cancer. The invention also relates to diagnostic methods and kits using these biomarkers and gene signatures.



5 **Gene signatures for classifying and grading lung cancer**

Field of the invention

[0001] The present invention relates to gene signatures that are indicative for the class and grade of lung cancer. The present invention also relates to methods of diagnosing classifying and grading a lung cancer tumor. The invention further
10 relates to arrays and computer readable media comprising such gene signatures.

Background of the Invention

[0002] Lung cancer is the most common cancer in the world, both in rates of incidence and in rates of mortality, and is most prevalent in Europe and North America. Lung cancer is often attributed to both genetic factors and exposure to
15 environmental factors (e.g., of radon gas, asbestos, and air pollution).

Accordingly, researchers have attempted to identify biomarkers for use in risk assessment, screening, diagnosis, prognosis, selection of therapy and monitoring therapy. *See, e.g.*, Yip et al. (WO 2004/061410), Semmens, et al. (WO 2005/098445), Sungwhan et al. (US 2007/0264659), Gold et al. (WO
20 2011/031344), and Birse (US 7,892,760).

[0003] Lung cancers are classified by histological type: non-small cell lung carcinoma (NSCLC), small cell lung carcinoma (SCLC), and carcinoid. Approximately 80% of all lung cancers are NCSLC, with SCLC and carcinoid accounting for approximately 17% and 1%, respectively. Approximately 2% of
25 lung cancers are unspecified.

[0004] Non-small cell lung carcinomas are grouped together based on similar prognosis and management and comprise three main sub-types: squamous cell lung carcinoma (SCC), adenocarcinoma and large cell lung carcinoma (LCLC). SCC account for approximately 25% of lung cancer incidents. Adenocarcinoma of the
30 lung accounts for approximately 40% of lung cancer incidents. LCLC is a heterogeneous group of undifferentiated malignant neoplasms originating from

transformed epithelial cells in the lung, which accounts for about 10% of lung cancer incidents.

[0005] NSCLC are staged according to the TNM system. The T category defines the primary tumor by size and whether it has spread into the surrounding tissue.

5 The N category identifies any lymph node involvement in and around the lungs. Finally, the M category indicates whether the cancer has metastasized. Under stage 1, the lung cancer is small and localized to a single area. Stage 2 and stage 3 lung cancers are larger, may have grown into surrounding tissues and may involve lymph nodes in and around the lungs. Stage 4 lung cancers have metastasized to
10 another region of the body.

[0006] Lung cancer is currently diagnosed by X-ray or computed tomography (CT) screening. X-ray analysis is typically performed if a patient reports symptoms that may suggest lung cancer and may reveal an obvious mass, widening of the mediastinum (suggestive of spread to lymph nodes there), atelectasis
15 (collapse or closure of alveoli), consolidation (pneumonia), or pleural effusion. CT imaging is subsequently performed to provide additional information about the type and extent of disease. After a suspicious mass is identified by X-ray and/or CT imaging, bronchoscopy or biopsy may be performed to analyze the suspicious tissue. Accordingly, the diagnostic tests for lung cancer usually require the disease
20 to have progressed to the point that lung function is moderately affected and tumor growth is visible. Thus, there is a need for a diagnostic test that can identify, classify and grade lung tumors in patients at early stages.

Summary of the invention

[0007] The present invention is directed to gene signatures for classifying,
25 diagnosing or grading lung cancer in an individual. A first aspect of the invention provides a method of classifying or grading a lung cancer tumor in an individual at risk for or having lung cancer. In some embodiments the method comprises classifying a test sample as stage 1 lung adenocarcinoma, stage 2 lung adenocarcinoma, stage 1 squamous cell carcinoma or stage 2 squamous cell
30 carcinoma. In some embodiments, the method comprises measuring the expression levels of at least 2 genes listed in Table 1 in a test sample; and applying one or more network-based methods, one or more machine-learning based

methods, or a combination of the foregoing methods to the expression levels to obtain a classification of the test sample as stage 1 lung adenocarcinoma, stage 2 lung adenocarcinoma, stage 1 squamous cell carcinoma or stage 2 squamous cell carcinoma. In some embodiments, a differential pattern of expression levels of
5 said at least 2 genes in the test sample classifies the lung cancer tumor as one of stage 1 lung adenocarcinoma, stage 2 adenocarcinoma, stage 1 squamous cell carcinoma or stage 2 squamous cell carcinoma.

[0008] In some embodiments, the differential pattern of expression levels is identified by a classifier based on a plurality of genes listed in Table 1, including
10 said at least two genes, said classifier having been trained by in silico analysis or one or more feature selection and classification algorithms. Optionally, the differential pattern of expression levels is identified by a classifier based on a plurality of genes listed in Table 1, including said at least two genes, said classifier having been trained by one or more algorithms selected from the group consisting
15 of dual ensemble, generalized simulated annealing, T-filter, CORG, CORG combined with support vector machine, dual bagging, single and pairs, forward learning, Laplacian based learning and learning method based on network perturbation amplitude. For example, the classifier may be trained with at least the data in the Gene Expression Omnibus datasets GSE2109, GSE10245, GSE18842
20 and GSE37745.

[0009] In some embodiments, the method comprises detecting the expression level of at least 2 of the genes listed in Table 1 in a test sample obtained from the individual; and comparing the expression level of the genes listed in Table 1 in the test sample to the expression level of the genes listed in Table 1 in a control
25 sample. In some embodiments, the method further comprises detecting the expression level of the genes listed in Table 1 in the control sample.

[0010] In some embodiments, the at least 2 genes are selected from the group consisting of: ZIC2, LOC100131262, CD83, EML1, PAIP1, NIPBL, CREB3L1, SLC37A1, and SFMBT2.

30 [0011] In some embodiments, the test sample or the control sample is selected from blood, serum, plasma, sputum, saliva, tissue, bronchia brushings, exhaled breath, and urine. Optionally, the tissue is lung tissue, such as tissue obtained by

biopsy from a tumor. In some embodiments, the control sample is lung tissue, such as tissue obtained by biopsy from healthy lung tissue. In some embodiments, the healthy lung tissue is obtained from the individual at risk for or having lung cancer. In other embodiments, the control sample is obtained from an individual
5 that does not have lung cancer.

[0012] In some embodiments, the expression level of the genes listed in Table 1 in the test sample and the expression level of the genes listed in Table 1 in the control sample are detected by measuring mRNA levels. Optionally, the expression level of the genes listed in Table 1 the test sample are detected by using
10 a human genome-wide array, a human lung tissue array or a custom array comprising polynucleotides of a plurality of genes in Table 1 and said at least 2 genes.

[0013] In some embodiments, the expression level of the genes listed in Table 1 in the test sample and the expression level of the genes listed in Table 1 in the control sample are detected by measuring the level of proteins encoded by the
15 genes.

[0014] In some embodiments, the expression level of the genes listed in Table 1 in the test sample and the expression level of the genes listed in Table 1 in the control sample are detected by measuring both mRNA levels and the level of
20 proteins encoded by the genes.

[0015] In some embodiments, the expression level of the genes listed in Table 1 in the test sample and the expression level of the genes listed in Table 1 in the control sample are compared by *in silico* analysis (e.g., network-based analysis or machine-learning methods).

[0016] A second aspect of the invention provides an array for use in classifying or grading a lung cancer tumor. In some embodiments, the array comprises polynucleotides immobilized on a solid surface that can hybridize to at least 10 lung cancer signature genes, wherein the lung cancer signature genes are selected from the group consisting of the genes listed in Table 1. Optionally, the array
30 comprises polynucleotides hybridizing to at least 2 lung cancer signature genes immobilized on a solid surface, wherein the lung cancer signature genes are

selected from the genes listed in Table 1. In some embodiments, the array is not a human genome-wide array.

[0017] A third aspect of the invention provides a panel for use in classifying or grading a lung cancer tumor. In some embodiments, the panel comprises
5 antibodies immobilized on a solid surface that bind to proteins encoded by at least 2 lung cancer signature genes, wherein the lung cancer signature genes are selected from the group consisting of the genes listed in Table 1.

[0018] A fourth aspect of the invention provides a computer readable medium for use in classifying or grading a lung cancer tumor. In some embodiments, the
10 computer readable medium comprises a lung cancer gene signature, wherein the gene signature comprises at least 2 genes selected from the genes listed in Table 1.

[0019] In some embodiments, the computer readable medium or computer program product comprises a classifier based on at least two genes listed in Table 1, said classifier having been trained by in silico analysis or one or more feature
15 selection and classification algorithms. Optionally, the classifier is trained by one or more algorithms selected from the group consisting of dual ensemble, generalized simulated annealing, T-filter, CORG, CORG combined with support vector machine, dual bagging, single and pairs, forward learning, Laplacian based learning and learning method based on network perturbation amplitude. The
20 classifier may be trained with at least the data in the Gene Expression Omnibus datasets GSE2109, GSE10245, GSE18842 and GSE37745. In some embodiments, the at least two genes are selected from the group consisting of ZIC2, LOC100131262, CD83, EML1, PAIP1, NIPBL, CREB3L1, SLC37A1, and SFMBT2.

[0020] A fifth aspect of the invention provides a device for classifying and grading a lung cancer tumor. In some embodiments, the device comprises means
25 for detecting the expression level of the genes listed in Table 1 in a test sample; means for correlating the expression level with a grade or classification of the tumor as stage 1 or stage 2 lung adenocarcinoma or stage 1 or stage 2 squamous cell carcinoma; and means for outputting the lung cancer tumor grade or
30 classification. Optionally, the device further comprises means for detecting the expression level of the genes listed in Table 1 in a control sample.

[0021] A sixth aspect of the invention provides a kit for classifying and grading a lung cancer tumor. In some embodiments, the kit comprises a set of reagents that detects expression levels of the genes listed in Table 1 in a test sample and instructions for using said kit for classifying and grading the lung cancer tumor. In
5 other embodiments, the kit is for assessing the prognosis of a lung cancer tumor in an individual. In such embodiments, the kit comprises a set of reagents that detects expression levels of the genes listed in Table 1 in a test sample from the individual and instructions for using said kit for determining the prognosis of the lung cancer tumor in said individual. In some embodiments, the set of reagents that detects
10 expression levels of the genes listed in Table 1 in the test sample may also be used to detect expression levels of the genes listed in Table 1 in a control sample.

[0022] An seventh aspect of the invention provides a method of diagnosing the stage of a lung cancer tumor in an individual or of assessing the prognosis of an individual with a lung cancer tumor. In some embodiments, the method comprises
15 a) measuring the expression level of at least 10 genes/biomarkers selected from the group consisting of the genes listed in Table 1 in a biological sample obtained from the individual; b) calculating a numerical biomarker score for the individual based on the expression levels of the biomarkers measured in step a); wherein the numerical biomarker score is predictive of the stage of lung cancer in the
20 individual. In some embodiments, the method comprises measuring the expression level of at least 10 genes/biomarkers selected from the group consisting of the genes listed in Table 1 in a biological sample obtained from the individual; calculating a numerical biomarker score for the individual based on the expression levels of the biomarkers measured in step a); wherein the numerical biomarker
25 score is predictive of the prognosis of the lung cancer in the individual.

[0023] In some embodiments, the biological sample is selected from blood, serum, plasma, sputum, saliva, tissue, bronchia brushings, exhaled breath, and urine. Optionally, the tissue is lung tissue, such as tissue obtained by biopsy from a tumor.

30 [0024] In some embodiments, the expression level of the genes listed in Table 1 in the biological sample is detected by measuring mRNA levels. Optionally, the expression level of the genes listed in Table 1 the test sample are detected by using

a human genome-wide array, a human lung tissue array or a custom array comprising polynucleotides of a plurality of genes in Table 1 and said at least 2 genes.

[0025] In some embodiments, the expression level of the genes listed in Table 1 in the biological sample is detected by measuring the level of proteins encoded by the genes.

[0026] In some embodiments, the numerical biomarker score is calculated by *in silico* analysis. The *in silico* analysis may be network based analysis or a machine-learning method.

[0027] In some embodiments, the biomarkers are proteins encoded by the genes selected from the group consisting of the genes listed in Table 1.

[0028] Particular embodiments of the invention are set forth in the following numbered paragraphs:

1. A method of classifying or grading a lung cancer tumor in an individual at risk for or having lung cancer comprising
 - (1) detecting the expression level of at least 2 genes listed in Table 1 in a test sample obtained from the individual; and
 - (2) comparing the expression level of said at least 2 genes in the test sample to the expression level of said at least 2 genes in a control sample,wherein,
if the expression level of said at least 2 genes is different in the test sample than in the control sample, then the lung cancer tumor can be classified and staged as stage 1 lung adenocarcinoma, stage 2 adenocarcinoma, stage 1 squamous cell carcinoma or stage 2 squamous cell carcinoma.
2. The method according to paragraph 1, wherein the method further comprises detecting the expression level of said at least 2 genes in the control sample.
3. The method according to paragraph 1 or 2, wherein the test sample is selected from blood, serum, plasma, sputum, saliva, tissue, bronchia brushings, exhaled breath, and urine.
4. The method according to paragraph 3, wherein the tissue is lung tissue.

5. The method according to paragraph 4, wherein the lung tissue is obtained by biopsy from a tumor.
6. The method according to any one of paragraphs 2-5, wherein the control sample is selected from blood, serum, plasma, sputum, saliva, tissue, bronchia
5 brushings, exhaled breath, and urine.
7. The method according to paragraph 6, wherein the tissue is lung tissue.
8. The method according to paragraph 7, wherein the lung tissue is obtained by biopsy from healthy lung tissue.
9. The method according to paragraph 8, wherein the healthy lung tissue is
10 obtained from the individual at risk for or having lung cancer.
10. The method according to paragraph 6, wherein the control sample is obtained from an individual that does not have lung cancer.
11. The method according to any one of paragraphs 1-10, wherein the expression level of said at least 2 genes in the test sample and the expression level
15 of said at least 2 genes in the control sample are detected by measuring mRNA levels.
12. The method according to paragraph 11, wherein the mRNA is measured by amplification, hybridization, mass spectroscopy, serial analysis of gene expression, or massive parallel signature sequencing.
- 20 13. The method according to paragraph 12, wherein the amplification is reverse transcription PCR, real time quantitative PCR, differential display or TaqMan PCR.
14. The method according to paragraph 12, wherein the hybridization is a dot blot, a slot blot, an RNase protection assay, microarray hybridization, or in situ
25 hybridization.
15. The method according to paragraph 12, wherein the mass spectroscopy is MALDI-TOF mass spectroscopy.
16. The method according to any one of paragraphs 1-10, wherein the expression level of said at least 2 genes in the test sample and the expression level
30 of said at least 2 genes in the control sample are detected by measuring the level of proteins encoded by the genes.

17. The method according to paragraph 16, wherein the protein level is measured using an antibody assay or by mass spectroscopy.
18. The method according to paragraph 17, wherein the antibody assay is selected from Western analysis, immunofluorescence, ELISA, and
5 immunohistochemistry.
19. The method according to any one of paragraphs 1-18, wherein the expression level of said at least 2 genes in the test sample and, optionally, the expression level of said at least 2 genes in the control sample are compared by *in silico* analysis.
- 10 20. The method according to paragraph 19, wherein the *in silico* analysis comprises using a classifier generated by one or more network-based methods or machine-learning based methods.
21. An array comprising polynucleotides hybridizing to at least 2 lung cancer signature genes immobilized on a solid surface, wherein the lung cancer signature
15 genes are selected from the group consisting of at least 2 genes listed in Table 1.
22. A panel comprising antibodies immobilized on a solid surface that bind to proteins encoded by at least 2 lung cancer signature genes, wherein the lung cancer signature genes are selected from the group consisting of the genes listed in Table 1.
- 20 23. A computer readable medium comprising a lung cancer gene signature, wherein the gene signature comprises at least 2 genes selected from the genes listed in Table 1.
35. A device for classifying and grading a lung cancer tumor, the device comprising: means for detecting the expression level of at least 2 genes listed in
25 Table 1 in a test sample; means for correlating the expression level with a classification of the tumor as stage 1 or stage 2 lung adenocarcinoma or stage 1 or stage 2 squamous cell carcinoma; and means for outputting the lung cancer tumor classification.
36. A kit for classifying and grading a lung cancer tumor, comprising one or
30 more reagents that detects expression levels of at least 2 genes listed in Table 1 in a test sample and instructions for using said kit for classifying and grading a lung cancer tumor.

37. A kit for assessing the prognosis of lung cancer in an individual, comprising a set of reagents that detects expression levels of at least 2 genes listed in Table 1 in a test sample from the individual and instructions for using said kit for determining the prognosis of lung cancer in said individual.

5 38. A method of diagnosing the stage of lung cancer in an individual, said method comprising the steps of:

a) measuring the expression level of at least 2 genes/biomarkers selected from the group consisting of the genes listed in Table 1 in a biological sample obtained from the individual;

10 b) calculating a numerical biomarker score for the individual based on the expression levels of the biomarkers measured in step a);

wherein the numerical biomarker score is predictive of the stage of lung cancer in the individual.

15 39. A method of assessing the prognosis of an individual with lung cancer, said method comprising the steps of:

a) measuring the expression level of at least 2 genes/biomarkers selected from the group consisting of the genes listed in Table 1 in a biological sample obtained from the individual;

20 b) calculating a numerical biomarker score for the individual based on the expression levels of the biomarkers measured in step a);

wherein the numerical biomarker score is predictive of the prognosis of the lung cancer in the individual.

25 40. A method of diagnosing, prognosing, classifying or grading lung cancer in a biological sample or an individual comprising measuring the expression levels of at least 2 genes listed in Table 1 in the biological sample or a test sample from the individual; and applying one or more network-based methods, one or more machine-learning based methods, or a combination of the foregoing methods to the expression levels to obtain a classification of the test sample as stage 1 lung adenocarcinoma, stage 2 lung adenocarcinoma, stage 1 squamous cell carcinoma or stage 2 squamous cell carcinoma.

30 41. The method according to paragraph 40, wherein a classifier or a previously established standard is used to classify a test sample as stage 1 lung

adenocarcinoma, stage 2 lung adenocarcinoma, stage 1 squamous cell carcinoma or stage 2 squamous cell carcinoma.

42. The method according to paragraph 41, wherein the classifier is obtained by training with a network-based method or a machine-learning based method using datasets obtained from subjects with stage 1 lung adenocarcinoma, stage 2 lung adenocarcinoma, stage 1 squamous cell carcinoma or stage 2 squamous cell carcinoma and datasets from subjects without lung cancer.

43. A method of classifying or grading a lung cancer tumor in an individual at risk for or having lung cancer comprising detecting the expression level of at least 2 genes listed in Table 1 in a test sample obtained from the individual; wherein a differential pattern of expression levels of said at least 2 genes in the test sample classifies the lung cancer tumor as one of stage 1 lung adenocarcinoma, stage 2 adenocarcinoma, stage 1 squamous cell carcinoma or stage 2 squamous cell carcinoma.

44. The method according to paragraph 43, wherein the differential pattern of expression levels is identified by a classifier based on a plurality of genes listed in Table 1, including said at least two genes, said classifier having been trained by *in silico* analysis or one or more feature selection and classification algorithms.

45. The method according to paragraph 43 or 44, wherein the differential pattern of expression levels is identified by a classifier based on a plurality of genes listed in Table 1, including said at least two genes, said classifier having been trained by one or more algorithms selected from the group consisting of dual ensemble, generalized simulated annealing, T-filter, CORG, CORG combined with support vector machine, dual bagging, single and pairs, forward learning, Laplacian based learning and learning method based on network perturbation amplitude.

46. The method according to any one of paragraphs 43-45, wherein said classifier having been trained with at least the data in the Gene Expression Omnibus datasets GSE2109, GSE10245, GSE18842 and GSE37745.

47. The method according to any one of paragraphs 43-46, wherein the method further comprises comparing the expression level of said at least 2 genes in the test sample and a control sample; or detecting the expression level of said at least 2

genes in the control sample and comparing the expression level of said at least 2 genes in the test sample and control sample, to identify the differential pattern.

48. The method according to any one of claims 43-47, wherein said at least 2 genes are selected from the group consisting of: ZIC2, LOC100131262, CD83,
5 EML1, PAIP1, NIPBL, CREB3L1, SLC37A1, and SFMBT2.

49. The method according to any one of paragraphs 43-48, wherein the expression level of said at least 2 genes in the test sample are detected by using a human genome-wide array, a human lung tissue array or a custom array comprising polynucleotides of a plurality of genes in Table 1 and said at least 2
10 genes.

50. The method according to any one of paragraphs 43-48, wherein the expression level of said at least 2 genes in the test sample are detected by measuring the level of proteins encoded by the genes.

51. An array comprising polynucleotides hybridizing to at least 2 lung cancer
15 signature genes immobilized on a solid surface, wherein the lung cancer signature genes are selected from the genes listed in Table 1 and said array is not a human genome-wide array.

52. A device comprising antibodies immobilized on a solid surface that bind to proteins encoded by at least 2 lung cancer signature genes, wherein the lung cancer
20 signature genes are selected from the group consisting of the genes listed in Table 1.

53. A computer readable medium or computer program product comprising a classifier based on at least two genes listed in Table 1, said classifier having been trained by *in silico* analysis or one or more feature selection and classification
25 algorithms.

54. The computer readable medium or computer program product according to paragraph 53, wherein said classifier is trained by one or more algorithms selected from the group consisting of dual ensemble, generalized simulated annealing, T-filter, CORG, CORG combined with support vector machine, dual bagging, single
30 and pairs, forward learning, Laplacian based learning and learning method based on network perturbation amplitude.

55. The computer readable medium or computer program product according to paragraph 53 or 54, wherein said classifier is trained with at least the data in the Gene Expression Omnibus datasets GSE2109, GSE10245, GSE18842 and GSE37745.

- 5 56. The computer readable medium or computer program product according to any one of paragraphs 53-55, wherein said at least two genes are selected from the group consisting of ZIC2, LOC100131262, CD83, EML1, PAIP1, NIPBL, CREB3L1, SLC37A1, and SFMBT2.

- 10 57. A kit for classifying and grading a lung cancer tumor or for assessing the prognosis of lung cancer in an individual, comprising one or more reagents that detects expression levels of at least 2 genes listed in Table 1 in a test sample and instructions for using said kit for classifying and grading a lung cancer tumor or for determining the prognosis of lung cancer in said individual.

Brief Description of the Drawings

- 15 [0029] **Figure 1** provides a features selection and classification algorithm(s) used for prediction of a gene signature.

Detailed Description of the invention

[0030] In order that the invention described herein may be fully understood, the following detailed description is set forth.

- 20 [0031] Unless defined otherwise, all technical and scientific terms used herein have the same meaning as those commonly understood by one of skill in the art to which this invention belongs. In case of conflict, the present specification, including definitions, will control. Although methods and materials similar or equivalent to those described herein can be used in the practice or testing of the present invention, suitable methods and materials are described below. The materials, methods and examples are illustrative only, and are not intended to be limiting. All publications, patents and other documents mentioned herein are incorporated by reference in their entirety.

- 25 [0032] Throughout this specification, the word “comprise” or variations such as “comprises” or “comprising” will be understood to imply the inclusion of a stated
- 30

integer or groups of integers but not the exclusion of any other integer or group of integers.

[0033] The term “antibody” refers to an immunoglobulin molecule capable of specific binding to a target, such as a carbohydrate, polynucleotide, lipid, polypeptide, etc., through at least one antigen recognition site, located in the variable region of the immunoglobulin molecule. As used herein, unless otherwise indicated by context, the term is intended to encompass not only intact polyclonal or monoclonal antibodies, but also engineered antibodies (e.g., chimeric, humanized and/or derivatized to alter effector functions, stability and other biological activities) and fragments thereof (such as Fab, Fab’, F(ab’)2, Fv), single chain (ScFv) and domain antibodies, including shark and camelid antibodies), and fusion proteins comprising an antibody portion, multivalent antibodies, multispecific antibodies (e.g., bispecific antibodies so long as they exhibit the desired biological activity) and antibody fragments as described herein, and any other modified configuration of the immunoglobulin molecule that comprises an antigen recognition site. An antibody includes an antibody of any class, such as IgG, IgA, or IgM (or sub class thereof), and the antibody need not be of any particular class. Depending on the antibody amino acid sequence of the constant domain of its heavy chains, immunoglobulins can be assigned to different classes. There are five major classes of immunoglobulins: IgA, IgD, IgE, IgG, and IgM, and several of these may be further divided into subclasses (isotypes), e.g., IgG1, IgG2, IgG3, IgG4, IgA1 and IgA2 in humans. The heavy chain constant domains that correspond to the different classes of immunoglobulins are called alpha, delta, epsilon, gamma, and mu, respectively. The subunit structures and three dimensional configurations of different classes of immunoglobulins are well known.

[0034] The term “array” refers to the arrangement of biomarker detection molecules, such as nucleic acid probes or antibodies, on a solid support that allows for high-throughput screening of a sample to detect the presence and/or quantity of a biomarker. Such arrays may be used, e.g., to evaluate the expression levels of several genes of interest in a single high-throughput reaction. The array may be a nucleic acid array, such as a nucleic acid microarray; a protein array, such as a

protein microarray; a peptide array, such as a peptide microarray; a tissue microarray, such as a tissue microarray or an antibody microarray, such as an antibody microarray. The solid substrate may be a microscopic bead, a glass slide, a plastic chip or a silicon chip.

5 [0035] The term “biomarker” refers to a characteristic whose presence, absence or level indicates a biological state. Typically, the properties of biomarkers indicate a normal process, a pathogenic process or a response to a pharmaceutical or therapeutic intervention. A biomarker can be a cell, a gene, a gene product, an enzyme, a hormone, a protein, a peptide, an antibody, a nucleic acid molecule, a
10 metabolite, a lipid, a free fatty acid, cholesterol or some other chemical compound. A biomarker can be a morphologic biomarker (for example, a histological change, DNA ploidy, malignancy-associated changes in the cell nucleus and premalignant lesions) or a genetic biomarker (for example, DNA mutations, DNA adducts and apoptotic index).

15 [0036] The term “classifying a lung cancer” refers to a method for determining the type of tumor from which a subject suffers. A subject can suffer from several different types of lung cancer, including but not limited to, adenocarcinoma, squamous cell carcinoma, large cell lung carcinoma, other non-small cell lung carcinomas, small cell lung carcinoma, carcinoid and unspecified lung cancer.
20 Accordingly, a lung cancer tumor may be classified as one of these types of lung cancer. A tumor may be classified based on histology, genetics or the presence, absence, alteration or levels of biomarkers. A lung cancer may be classified based on the lung cancer gene signature. The lung cancer may be classified as either adenocarcinoma or squamous cell carcinoma.

25 [0037] As used herein, the term “computer program” refers to a sequence of instructions, written to perform a specified task within a computer. For example, a computer program product is described, the product comprising computer-readable instructions that, when executed in a computerized system comprising at least one
30 processor, cause the processor to carry out one or more steps of any of the methods described above. In another example, a computerized system is described, the system comprising a processor configured with non-transitory computer-readable

instructions that, when executed, cause the processor to carry out any of the methods described herein. The computer program product and the computerized methods described herein may be implemented in a computerized system having one or more computing devices, each including one or more processors.

- 5 Generally, the computerized systems described herein may comprise one or more engines, which include a processor or devices, such as a computer, microprocessor, logic device or other device or processor that is configured with hardware, firmware, and software to carry out one or more of the computerized methods described herein. Any one or more of these engines may be physically separable
10 from any one or more other engines, or may include multiple physically separable components, such as separate processors on common or different circuit boards. The computer systems of the present invention comprises means for implementing the methods and its various embodiments as described herein. The computerized system described herein may include a distributed computerized system having one
15 or more processors and engines that communicate through a network interface. Such an implementation may be appropriate for distributed computing over multiple communication systems.

- [0038] The term “computer readable medium” refers to a medium capable of storing data, such that the data may be accessed by a computer. Such a medium
20 may take many forms, including but not limited to, non-volatile media and volatile media. Non-volatile media include, for example, optical, magnetic, or opto-magnetic disks, or integrated circuit memory, such as flash memory. Volatile media include, for example, dynamic random access memory (DRAM), which typically constitutes the main memory. Common forms of computer-readable
25 media include, for example, a floppy disk, a flexible disk, hard disk, magnetic tape, magnetic cards, magnetic ink characters, magnetic drums, any other magnetic medium, a CD-ROM, DVD, any other optical medium, punch cards, paper tape, any other physical medium with patterns of holes, a RAM, a PROM, an EPROM or EEPROM (electronically erasable programmable read-only memory), a
30 FLASH-EEPROM, barcodes, semiconductors, microchips and any other memory chip or cartridge.

[0039] The term “control sample” refers to a sample against which a test sample is compared in order to diagnose, prognose, classify or grade the test sample. A control sample may be healthy tissue or may be a well-characterized tumor sample, including but not limited to, stage 1 adenocarcinoma, stage 2 adenocarcinoma, stage 1 squamous cell carcinoma, or stage 2 squamous cell carcinoma. A control sample can be analyzed concurrently with or separately from the test sample, including before or after analyzing the test sample. The data from the analysis of a control sample may be stored, e.g., in a computer readable medium or in a manual, for comparison against test samples analyzed in the future or as data for training network-based or machine-learning methods. A control sample may be developed as a medical standard for comparison. For example, analysis of control samples has developed medical standards for normal fed and fasted blood glucose levels; normal, at risk, and hypertensive blood pressures, and normal resting heart rates. As used herein, the term “control sample” includes samples that provided a medical standard. Accordingly, a test sample may be compared against a medical standard generated from control samples. For example, expression of a variant or mutated form of a gene may be indicative of a change medical condition. Alternatively, a change in expression level of a gene may be indicative of a change in medical condition. A control sample may be lung tissue, such as tissue obtained by biopsy from a healthy portion of a lung (e.g., distant from a suspected tumor), or some other sample. For example, a control sample may be blood, blood cells, serum, plasma, sputum, saliva, tissue, bronchial washing, bronchial aspirates, bronchia brushings, exhaled breath, lymph fluid, and urine. Tissue specimens, such as those obtained by biopsy, may be fixed (e.g., formaldehyde-fixed paraffin-embedded (FFPE)). The control sample may be obtained from a tissue bank. The control sample may also be obtained from a cadaver or an organ donor.

[0040] The terms “differential pattern of expression” and “differential expression” are used interchangeably herein and refer to a difference in an activity measurement (e.g., the variability or difference of genetic expression) of a biological entity under different conditions. For example, one condition may refer to an experimental treatment (such as exposure to a potentially carcinogenic agent), and another condition may refer to a control treatment (such as a null

treatment). In an example, a fold-change is a number describing how much a measurement at a node (or biological entity) changes from an initial value to a final value between control data and treatment data, or between two sets of data representing different treatment conditions. The fold-change number may represent the logarithm of the fold-change of the activity of the biological entity between the two conditions. A confidence interval for the significance of the fold-change number may also be assessed.

[0041] The terms “gene signature” and “genetic signature” are used interchangeably herein and refer to a group of genes expressed in a cell, whose combined expression pattern may be indicative of, e.g., a normal state, an at-risk state, a diseased state (e.g., stage 1 lung adenocarcinoma, stage 2 lung adenocarcinoma, stage 1 squamous cell carcinoma or stage 2 squamous cell carcinoma), a treated state or a recovery state. A gene signature may be characterized by which genes are expressed and/or at what level each gene is expressed. Gene signatures are particularly useful in diagnosing, prognosing, classifying or grading complex diseases states, which result from the combination of several genetic and environmental factors. The gene signatures disclosed herein may be used, e.g., for the diagnosis, prognosis, classification or grading of lung cancer tumors in an individual. The gene signature may be unique to the class and grade of the tumor.

[0042] The term “grading a lung cancer” refers to a method for determining the grade of tumor from which a subject suffers. A subject can suffer from several different grades of lung cancer, which reflect the severity and invasiveness of the lung cancer. For example, stage 0 refers to a carcinoma in situ; stage 1 refers to cancers that are localized to one part of the body; stage 2 refers to cancers that are locally advanced; stage 3 refers to cancers that are further advanced locally (e.g., as evidenced by increased lymph node involvement) and stage 4 refers to cancers that have metastasized. Lung cancers are typically staged according to the TNM system. The T category defines the primary tumor by size and whether it has spread into the surrounding tissue. The N category identifies any lymph node involvement in and around the lungs. Finally, the M category indicates whether the cancer has metastasized. A tumor may be graded based on histology, genetics

or the presence, absence, or levels of biomarkers. A lung cancer may be graded based on its gene signature. The lung cancer may be graded as stage 1 adenocarcinoma, stage 2 adenocarcinoma, stage 1 squamous cell carcinoma or stage 2 squamous cell carcinoma.

- 5 **[0043]** The term “*in silico* analysis” refers to analysis performed on a computer or via computer simulation. Gene signature analysis involves detection of gene expression based on identity and expression level for a multitude of genes. *In silico* analysis may apply one or more network-based methods, one or more machine-learning based methods, or a combination of the foregoing methods to the
- 10 expression levels to obtain a classification of the test sample, e.g., as stage 1 lung adenocarcinoma, stage 2 lung adenocarcinoma, stage 1 squamous cell carcinoma or stage 2 squamous cell carcinoma. Comparisons between expression levels from test samples and control samples may require computer analysis to determine the degree and significance of any changes observed. *See, e.g.*, U.S. Provisional
- 15 Patent Application entitled “Systems and Methods relating to Network-based Biomarker Signatures,” filed concurrently with the instant application, incorporated herein by reference in its entirety and having the attorney docket no. 106500-0022-001; U.S. Provisional Patent Application entitled “Systems and Methods for Generating Biomarker Signatures,” filed concurrently with the instant
- 20 application, incorporated herein by reference in its entirety and having the attorney docket no. 106500-0028-001; U.S. Provisional Patent Application entitled “Systems and Methods for Generating Biomarker Signatures with Integrated Bias Correction and Class Prediction,” filed concurrently with the instant application, incorporated herein by reference in its entirety and having the attorney docket no.
- 25 106500-0032-001; and U.S. Provisional Patent Application entitled “Systems and Methods for Generating Biomarker Signatures with Integrated Dual Ensemble and Simulated Annealing Techniques,” filed concurrently with the instant application, incorporated herein by reference in its entirety and having the attorney docket no. 106500-0031-001.
- 30 **[0044]** The term “individual” refers to a vertebrate, preferably a mammal. The mammal can be, without limitation, a mouse, a rat, a cat, a dog, a horse, a pig, a cow, a non-human primate or a human.

[0045] The term “individual at risk for” lung cancer” refers to an individual who is predisposed to lung cancer. Predisposition to lung cancer may be due to one or more genetic or environmental factors. For example, an individual related to a lung cancer patient is 2.4 times more likely to get lung cancer than an individual who is not related to a lung cancer patient. Further, exposure to environmental factors such as radon gas, asbestos, tobacco smoke, and air pollution can increase the risk for lung cancer and predispose an individual to lung cancer.

[0046] The term “individual having lung cancer” or “individual suffering from lung cancer” refers to an individual experiencing uncontrolled cell growth in the tissues of the lung. Lung cancers typically form solid tumors, which can be observed on a chest X-ray or by a CT scan.

[0047] The term “MALDI-TOF” refers matrix-assisted laser desorption/ionization time of flight mass spectroscopy. Matrix-assisted laser desorption/ionization (MALDI) is a two step process that uses laser-triggered desorption of protonated and deprotonated matrix materials to protonate or deprotonate analyte molecules (e.g., DNA, RNA, and proteins). Time-of-flight (TOF) mass spectrometry refers to a method in which an ion’s mass-to-charge ratio is determined by measuring the time that it takes an ionized particle to reach a detector at a known distance.

[0048] The term “ network-based analysis” refers to an approach to identify biomarkers that is based on the properties of groups of functionally interrelated genes that form a network in a biological system, instead of treating individual genes in the biological system a priori as completely independent and identical.

[0049] The term “machine learning methods” refers to methods that allow a machine, such as a programmable computer, to improve its performance at a certain predictive task that is based on the known properties of examples or training data. Machine learning methods include, without limitation, support vector machines (SVMs), network-based SVMs, ensemble classifiers, neural network-based classifiers, logistic regression classifiers, decision tree-based classifiers, classifiers employing a linear discriminant analysis technique, a random-forest analysis technique, or both.

[0050] The term “numerical biomarker score” refers to a number that is representative of the result(s) of one or more of the network-based analysis or machine learning methods.

[0051] The term “polynucleotide hybridizing to” refers to a polynucleotide molecule that binds to a target nucleic acid molecule through complementary base pair sequencing. Hybridization typically requires two nucleic acids that contain complementary sequences, although depending on the stringency of the hybridization, mismatches between bases are possible. The appropriate stringency for hybridizing nucleic acids depends on the length of the nucleic acids and the degree of complementation, variables well known in the art. Exemplary high stringent hybridization conditions are equivalent to about 20-27 °C below the melting temperature (T_m) of the DNA duplex formed in about 1 M salt. Many equivalent procedures exist and several popular molecular cloning manuals describe suitable conditions for highly stringent hybridization and, furthermore, provide formulas for calculating the length of hybrids expected to be stable under these conditions (see, e.g., Current Protocols in Molecular Biology, John Wiley & Sons, N.Y. (1989), 6.3.1 6 or 13.3.6; or pages 9.47-9.57 of Sambrook, et al. (1989) Molecular Cloning, 2nd ed., Cold Spring Harbor Press). “High stringency” refers to hybridization and/or washing conditions at 68 °C in 0.2 x SSC, at 42 °C in 50 % formamide, 4 x SSC, or under conditions that afford levels of hybridization equivalent to those observed under either of these two conditions. The greater the degree of similarity or homology between two nucleotide sequences, the greater the value of T_m , for hybrids of nucleic acids having those sequences. The relative stability (corresponding to higher T_m) of nucleic acid hybridizations decreases in the following order: RNA:RNA, DNA:RNA, DNA:DNA.

[0052] The terms “protein,” “polypeptide” and “peptide” are used interchangeably and indicate at least one molecular chain of amino acids linked through covalent or non-covalent bonds. The terms do not refer to a specific length of the molecular chain. Peptides, oligopeptides and proteins are included within the definition of “polypeptide”. The terms include post-translational modifications of the molecule, e.g., phosphorylation, glycosylation and acetylation.

The terms also include protein fragments, fusion proteins, mutant proteins and variant proteins.

[0053] The term “SELDI-TOF” refers surface-enhanced laser desorption/ionization time of flight mass spectroscopy. Surface-enhanced laser desorption/ionization (SELDI) is a variant of MALDI that uses a target with a biochemical affinity for the analyte. Time-of-flight (TOF) mass spectrometry refers to a method in which an ion’s mass-to-charge ratio is determined by measuring the time that it takes an ionized particle to reach a detector at a known distance.

10 [0054] The term “test sample” refers to a sample obtained from an individual at risk for, having or suffering from lung cancer. A test sample may be any sample suspected of containing or exhibiting a biomarker. The test sample is analyzed and compared to a control sample, including medical standards developed from control samples, to diagnose, prognose, classify or grade lung cancer in the individual. A
15 test sample may be obtained from lung tissue, such as tissue obtained by biopsy from a tumor, or other biological tissue. For example, a test sample may be blood, blood cells, serum, plasma, sputum, saliva, tissue, bronchial washing, bronchial aspirates, bronchia brushings, exhaled breath, lymph fluids, and urine. Tissue specimens, such as those obtained by biopsy, may be fixed (e.g., formaldehyde-
20 fixed paraffin-embedded (FFPE)).

[0055] As used herein, to “train” a data set means to generate a classifier that can accurately predict classifications of a set of test samples. For example, a training data set includes a set of samples, and each sample may correspond to a measurement from a different patient. A machine learning technique is applied to
25 the training data set to generate a “classifier,” which corresponds to a way of assigning each sample in the training data set to a category (such as “disease positive” or “disease free”). In addition to the training data set, a training class set is known. The training class set includes a known category assigned to each sample (or person). The categories predicted by the classifier are compared to the
30 known categories. If the predicted categories mostly match the known categories, the classifier has performed well. However, if there are substantial differences between the predicted categories and the known categories, the parameters of the

machine learning technique may be updated, and the updated machine learning technique is applied. These steps are repeated until the performance of a classifier exceeds a threshold, and the final classifier is provided. The final classifier may then be applied to a test data set. The test data set may correspond to measured
5 samples from different patients, but the patients in the test data set may have unknown categories (disease states). Thus, applying the final classifier to the test data set thus allows for prediction of the disease states of the patients.

Gene Signatures

[0056] One aspect of the invention provides gene signatures useful for
10 diagnosing, prognosing, classifying or grading a lung cancer tumor. In some embodiments, the gene signature comprises at least 2 genes selected from the genes listed in Table 1. In some embodiments, the gene signature comprises at least 2, at least 3, at least 4, at least 5, at least 10, at least 15, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at
15 least 65, at least 70, at least 75, at least 80, least about 85 or at least 87 genes selected from the genes listed in Table 1. In some embodiments, the gene signature comprises each of the genes listed in Table 1. Optionally, said at least 2, at least 3, at least 4, or at least 5 genes are selected from the group consisting of: ZIC2, LOC100131262, CD83, EML1, PAIP1, NIPBL, CREB3L1, SLC37A1, and
20 SFMBT2, which are the genes that appear in 4 of the 5 lists generated in Example 1.

[0057] In some embodiments, at least 2, at least 3, at least 4, at least 5, at least 10, at least 15, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, or at least 70 of the genes selected
25 from the genes listed in Table 1 have increased expression compared to a control sample. In some embodiments, at least 2, at least 3, at least 4, at least 5, at least 10, at least 15, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, or at least 70 of the genes selected from the genes listed in Table 1 have decreased expression compared to a control
30 sample.

[0058] In some embodiments, the gene signature includes a degree of up-regulation of a subset of genes in the gene signature compared to the control

sample. For example, each up-regulated gene in the gene signature may, independently, be up-regulated at least 1.5-fold, at least 2-fold, at least 2.5-fold, at least 3-fold, at least 3.5-fold, at least 4-fold, at least 4.5-fold, at least 5-fold, at least 6-fold, at least 7-fold, at least 8-fold, at least 9-fold, at least 10-fold, at least 100-fold, at least 1,000-fold or more compared to the control sample. Similarly, in some embodiments, the gene signature includes a degree of down-regulation of a subset of genes in the gene signature compared to the control sample. For example, each down-regulated gene in the gene signature may, independently, be down-regulated at least 1.5-fold, at least 2-fold, at least 2.5-fold, at least 3-fold, at least 3.5-fold, at least 4-fold, at least 4.5-fold, at least 5-fold, at least 6-fold, at least 7-fold, at least 8-fold, at least 9-fold, at least 10-fold, at least 100-fold, at least 1,000-fold or more compared to the control sample.

[0059] The present invention encompasses the following gene signatures:

- i. A and B;
- 15 ii. A, B, and C;
- iii. A, B, C, and D;
- iv. A, B, C, D, and E;
- v. A, B, C, D, E, and F;
- vi. A, B, C, D, E, F, and G;
- 20 vii. A, B, C, D, E, F, G, and H;
- viii. A, B, C, D, E, F, G, H, and I;
- ix. A, B, C, D, E, F, G, H, I, and J;
- x. A, B, C, D, E, F, G, H, I, J, and K;
- xi. A, B, C, D, E, F, G, H, I, J, K, and L;
- 25 xii. A, B, C, D, E, F, G, H, I, J, K, L and M;
- xiii. A, B, C, D, E, F, G, H, I, J, K, L, M, and N;
- xiv. A, B, C, D, E, F, G, H, I, J, K, L, M, N, and O;
- xv. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, and P;
- xvi. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, and Q;
- 30 xvii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, and R;
- xviii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, and S;
- xix. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, and T;

- xx. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, and U;
- xxi. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, and V;
- xxii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, and W;
- xxiii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, and X;
- 5 xxiv. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, and Y;
- xxv. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, and Z;
- xxvi. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, and AA;
- 10 xxvii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, and BB;
- xxviii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB and AC;
- 15 xxix. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, and AD;
- xxx. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD and AE;
- xxxi. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE and AF;
- 20 xxxii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF and AG;
- xxxiii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG and AH;
- 25 xxxiv. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH and AI;
- xxxv. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI and AJ;
- xxxvi. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ and AK;
- 30 xxxvii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK and AL;

- xxxviii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL and AM;
- xxxix. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM and AN;
- 5 xl. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN and AO;
- xli. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO and AP;
- 10 xlii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP and AQ;
- xliii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP, AQ and AR;
- 15 xliv. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP, AQ, AR and AS;
- xlv. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP, AQ, AR, AS and AT;
- 20 xlvi. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP, AQ, AR, AS, AT and AU;
- 25 xlvii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP, AQ, AR, AS, AT, AU and AV;
- xlviii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP, AQ, AR, AS, AT, AU, AV and AW;
- 30

- xlix. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP, AQ, AR, AS, AT, AU, AV, AW and AX;
- 5 i. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP, AQ, AR, AS, AT, AU, AV, AW, AX and AY;
- ii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP, AQ, AR, AS, AT, AU, AV, AW, AX, AY and AZ;
- 10 iii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP, AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ and BA;
- liii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP, AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA and BB;
- 15 liv. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP, AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB and BC;
- lv. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP, AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC and BD;
- 20 lvi. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP, AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD and BE;
- 25 lvii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP, AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE and BF;
- lviii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP, AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF and BG;
- 30

- lix. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,
AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF,
BG and BH;
- 5 lx. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,
AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF,
BG, BH and BI;
- 10 lxi. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,
AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF,
BG, BH, BI and BJ;
- 15 lxii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,
AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF,
BG, BH, BI, BJ and BK;
- 20 lxiii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,
AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF,
BG, BH, BI, BJ, BK and BL;
- 25 lxiv. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,
AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF,
BG, BH, BI, BJ, BK, BL and BM;
- 30 lxv. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,
AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF,
BG, BH, BI, BJ, BK, BL, BM and BN;
- 30 lxvi. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,
AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF,
BG, BH, BI, BJ, BK, BL, BM, BN and BO;

- lxvii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,
AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF,
BG, BH, BI, BJ, BK, BL, BM, BN, BO and BP;
- 5 lxviii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,
AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF,
BG, BH, BI, BJ, BK, BL, BM, BN, BO, BP and BQ;
- 10 lxix. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,
AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF,
BG, BH, BI, BJ, BK, BL, BM, BN, BO, BP, BQ and BR;
- 15 lxx. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,
AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF,
BG, BH, BI, BJ, BK, BL, BM, BN, BO, BP, BQ, BR and BS;
- 20 lxxi. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,
AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF,
BG, BH, BI, BJ, BK, BL, BM, BN, BO, BP, BQ, BR, BS and BT;
- 25 lxxii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,
AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF,
BG, BH, BI, BJ, BK, BL, BM, BN, BO, BP, BQ, BR, BS, BT and BU;
- 30 lxxiii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,
AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF,
BG, BH, BI, BJ, BK, BL, BM, BN, BO, BP, BQ, BR, BS, BT, BU and BV;
- lxxiv. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,
AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF,

BG, BH, BI, BJ, BK, BL, BM, BN, BO, BP, BQ, BR, BS, BT, BU, BV and BW;

lxxv. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,
5 AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF,
BG, BH, BI, BJ, BK, BL, BM, BN, BO, BP, BQ, BR, BS, BT, BU, BV,
BW and BX;

lxxvi. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,
10 AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF,
BG, BH, BI, BJ, BK, BL, BM, BN, BO, BP, BQ, BR, BS, BT, BU, BV,
BW, BX and BY;

lxxvii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,
15 AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF,
BG, BH, BI, BJ, BK, BL, BM, BN, BO, BP, BQ, BR, BS, BT, BU, BV,
BW, BX, BY and BZ;

lxxviii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,
20 AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF,
BG, BH, BI, BJ, BK, BL, BM, BN, BO, BP, BQ, BR, BS, BT, BU, BV,
BW, BX, BY, BZ and CA;

lxxix. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,
25 AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF,
BG, BH, BI, BJ, BK, BL, BM, BN, BO, BP, BQ, BR, BS, BT, BU, BV,
BW, BX, BY, BZ, CA and CB;

lxxx. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,
30 AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF,
BG, BH, BI, BJ, BK, BL, BM, BN, BO, BP, BQ, BR, BS, BT, BU, BV,
BW, BX, BY, BZ, CA, CB and CC;

- lxxxix. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,
AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF,
BG, BH, BI, BJ, BK, BL, BM, BN, BO, BP, BQ, BR, BS, BT, BU, BV,
5 BW, BX, BY, BZ, CA, CB, CC and CD;
- lxxxii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,
AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF,
BG, BH, BI, BJ, BK, BL, BM, BN, BO, BP, BQ, BR, BS, BT, BU, BV,
10 BW, BX, BY, BZ, CA, CB, CC, CD and CE;
- lxxxiii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,
AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF,
BG, BH, BI, BJ, BK, BL, BM, BN, BO, BP, BQ, BR, BS, BT, BU, BV,
15 BW, BX, BY, BZ, CA, CB, CC, CD, CE and CF; and
- lxxxiv. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,
AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF,
BG, BH, BI, BJ, BK, BL, BM, BN, BO, BP, BQ, BR, BS, BT, BU, BV,
20 BW, BX, BY, BZ, CA, CB, CC, CD, CE, CF and CG;
- lxxxv. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,
AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF,
BG, BH, BI, BJ, BK, BL, BM, BN, BO, BP, BQ, BR, BS, BT, BU, BV,
25 BW, BX, BY, BZ, CA, CB, CC, CD, CE, CF, CG and CH;
- lxxxvi. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,
AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF,
BG, BH, BI, BJ, BK, BL, BM, BN, BO, BP, BQ, BR, BS, BT, BU, BV,
30 BW, BX, BY, BZ, CA, CB, CC, CD, CE, CF, CG, CH and CI;
- lxxxvii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z,
AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP,

AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF, BG, BH, BI, BJ, BK, BL, BM, BN, BO, BP, BQ, BR, BS, BT, BU, BV, BW, BX, BY, BZ, CA, CB, CC, CD, CE, CF, CG, CH, CI and CJ;

wherein each of A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP, AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF, BG, BH, BI, BJ, BK, BL, BM, BN, BO, BP, BQ, BR, BS, BT, BU, BV, BW, BX, BY, BZ, CA, CB, CC, CD, CE, CF, CG, CH, CI and CJ are independently selected from the genes listed in Table 1 and each of A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO, AP, AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA, BB, BC, BD, BE, BF, BG, BH, BI, BJ, BK, BL, BM, BN, BO, BP, BQ, BR, BS, BT, BU, BV, BW, BX, BY, BZ, CA, CB, CC, CD, CE, CF, CG, CH, CI and CJ are different.

15 Methods of Using Biomarkers and Gene Signatures

[0060] The biomarkers and gene signatures of the invention may be used in methods of diagnosing, prognosing, classifying or grading lung cancer in a biological sample or an individual. The invention encompasses a method for classifying a test sample as stage 1 lung adenocarcinoma, stage 2 lung adenocarcinoma, stage 1 squamous cell carcinoma or stage 2 squamous cell carcinoma, the method comprising: measuring the expression levels of at least 2 genes listed in Table 1 in a test sample; and applying one or more network-based methods, one or more machine-learning based methods, or a combination of the foregoing methods to the expression levels to obtain a classification of the test sample as stage 1 lung adenocarcinoma, stage 2 lung adenocarcinoma, stage 1 squamous cell carcinoma or stage 2 squamous cell carcinoma. In some embodiments the expression levels of at least 2, at least 3, at least 4, at least 5, at least 10, at least 15, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, least about 85 or at least 87 genes selected from the genes listed in Table 1 are measured. In some embodiments, a differential pattern of expression levels of said at least 2 genes in the test sample classifies the lung cancer tumor as one of

stage 1 lung adenocarcinoma, stage 2 adenocarcinoma, stage 1 squamous cell carcinoma or stage 2 squamous cell carcinoma.

[0061] In one embodiment, the methods of the invention can be used to identify a gene signature and a classifier (e.g., a gene-signature-based classifier) that can distinguish datasets obtained from various classes and stages of lung cancer (e.g. stage 1 lung adenocarcinoma, stage 2 lung adenocarcinoma, stage 1 squamous cell carcinoma or stage 2 squamous cell carcinoma). In some methods of the invention, control data is not collected or used. Instead, a classifier or a previously established standard may be used to determine whether a test sample is a lung cancer sample or the class and stage of lung cancer from which the test sample is obtained. For example, a classifier that is obtained by training with network-based or machine-learning methods using datasets obtained from subjects with various classes and stages of lung cancer (e.g. stage 1 lung adenocarcinoma, stage 2 lung adenocarcinoma, stage 1 squamous cell carcinoma or stage 2 squamous cell carcinoma) and datasets from subjects without lung cancer, can be used. Alternatively, one or more numerical scores (e.g., average fold change or rank absolute) generated by the algorithms described herein may be used as a previously established standard. The levels of expression of one or more of the genes listed in Table 1 in a test sample may be compared to the previously established standard, and the comparison may be used to classify the test sample as a lung cancer sample or a normal sample. In some embodiments, the comparison may be used to classify the test sample as stage 1 lung adenocarcinoma, stage 2 lung adenocarcinoma, stage 1 squamous cell carcinoma or stage 2 squamous cell carcinoma.

[0062] In one embodiment, the invention provides a method of diagnosing, prognosing, classifying or grading lung cancer in a biological sample, wherein the method comprises determining the properties (for example, absence, presence or expression level) of one or more genes listed in Table 1 in the biological sample; and applying *in silico* analysis with a classifier obtained from a network-based method, a machine-learning based method, or a combination of the foregoing methods. The classifier can be obtained from the network-based method, a machine-learning based method, or a combination of the foregoing methods by

training with datasets obtained from subjects with stage 1 lung adenocarcinoma, stage 2 lung adenocarcinoma, stage 1 squamous cell carcinoma stage 2 squamous cell carcinoma, healthy subjects, or a combination of two or more of the foregoing. In another embodiment, such a classifier may be linked to a specific prognosis of the lung cancer in the individual who provided the biological sample. In a further embodiment, the classifier may indicate that the lung cancer in the individual who provided the biological sample is stage 1 lung adenocarcinoma, stage 2 lung adenocarcinoma, stage 1 squamous cell carcinoma or stage 2 squamous cell carcinoma. The classifier may also indicate that a particular treatment regimen should be used to treat the individual who provided the biological sample.

[0063] In one embodiment, the methods of the invention comprise obtaining a test sample (such as a lung biopsy) from an individual, determining the absence, presence or expression level of one or more of the genes listed in Table 1 in the test sample, comparing said absence, presence or expression level to the absence, presence or expression level of the same gene(s) in a control sample, and selecting a lung cancer treatment regimen based on the comparison. In a further embodiment, the invention provides a method for monitoring the progress of a lung cancer treatment in an individual, said method comprising determining at suitable time intervals before, during, or after therapy (for example, at different time points during the treatment) in a sample taken from said individual differential expression of a panel of at least 2 genes selected from the genes listed in Table 1.

[0064] In one embodiment, the invention encompasses a method that comprises collecting data on the properties of one or more genes in the gene signature without generating a gene signature. For example, the method of the invention comprises obtaining a test sample from an individual, and detecting the absence, presence or the expression level of one or more of the genes listed in Table 1 in the sample. In one embodiment, the invention encompasses a method that comprises using data on the properties of one or more genes in a gene signature that are already collected as training data to generate an improved gene signature using one or more network-based methods, one or more machine learning methods, or a combination of the foregoing methods. In one embodiment, the invention encompasses a method that comprises collecting data on the properties of one or

more genes in a biological system which is included in a gene signature, and using the data to predict a classification of the state of the biological system associated with the collected data.

[0065] In some embodiments, the method comprises detecting the expression level of at least 2 of the genes listed in Table 1 in a test sample obtained from the individual; and comparing the expression level of the genes listed in Table 1 in the test sample to the expression level of the genes listed in Table 1 in a control sample. In some embodiments, if the expression level of the genes listed in Table 1 is different in the test sample than in the control sample, then the individual suffers from lung cancer. In some embodiments, the lung cancer subtype and stage may be determined. In some embodiments, the lung cancer is classified as stage 1 lung adenocarcinoma, stage 2 lung adenocarcinoma, stage 1 squamous cell carcinoma or stage 2 squamous cell carcinoma. In some embodiments, the method further comprises detecting the expression level of the genes listed in Table 1 in the control sample. In some embodiments, the expression levels of at least 2, at least 3, at least 4, at least 5, at least 10, at least 15, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, at least 84 genes or all genes listed in Table 1 are detected.

[0066] In some embodiments, the expression level of the genes listed in Table 1 in the test sample and the expression level of the genes listed in Table 1 in the control non-tumor biological sample are compared by *in silico* analysis. The *in silico* analysis may be network based analysis or a machine-learning method.

[0067] In some embodiments, the test sample is selected from blood, serum, plasma, sputum, saliva, tissue, bronchia brushings, exhaled breath, and urine. Optionally, the tissue is lung tissue, such as tissue obtained by biopsy from a tumor.

[0068] In some embodiments, the control sample is selected from blood, serum, plasma, sputum, saliva, tissue, bronchia brushings, exhaled breath, and urine. In some embodiments, the tissue is lung tissue, such as tissue obtained by biopsy from healthy lung tissue. In some embodiments, the healthy lung tissue is obtained

from the individual at risk for or having lung cancer. In other embodiments, the control sample is obtained from an individual that does not have lung cancer.

[0069] In some embodiments, the expression level of the genes listed in Table 1 in the test sample and the expression level of the genes listed in Table 1 in the control sample are detected by measuring mRNA levels. For example, mRNA level is measured by amplification, hybridization, mass spectroscopy, serial analysis of gene expression, or massive parallel signature sequencing. Optionally, the amplification is reverse transcription PCR, real time quantitative PCR, differential display or TaqMan PCR. In some embodiments, the hybridization is a dot blot, a slot blot, an RNase protection assay, microarray hybridization, or in situ hybridization. The mass spectroscopy may be MALDI-TOF mass spectroscopy. In some embodiments, the expression level of the genes listed in Table 1 in the test sample are detected by using a human genome-wide array, a human lung tissue array or a custom array comprising polynucleotides of a plurality of genes in Table 1.

[0070] In some embodiments, the expression level of the genes listed in Table 1 in the test sample and the expression level of the genes listed in Table 1 in the control sample are detected by measuring the level of proteins encoded by the genes. Optionally, the protein level is measured using an antibody assay or by mass spectroscopy. In some embodiments, the antibody assay is selected from Western analysis, immunofluorescence, ELISA, and immunohistochemistry. The mass spectroscopy may be MALDI-TOF mass spectroscopy or SELDI-TOF mass spectroscopy.

[0071] In some embodiments, the expression level of the genes listed in Table 1 in the test sample and the expression level of the genes listed in Table 1 in the control sample are detected by measuring both mRNA levels and the level of proteins encoded by the genes. In some embodiments, expression levels are measured using the amplification, hybridization, mass spectroscopy, serial analysis of gene expression, massive parallel signature sequencing, and antibody assays discussed above.

Methods of Biomarker Detection, Arrays and Panels

[0072] Detection of the nucleic acid and/or protein biomarkers described herein in a test sample or a control sample may be performed in a variety of ways.

5 [0073] In one aspect, the methods of the invention rely on the detection of the presence or absence of biomarker genes and/or biomarker gene expression, or the qualitative or quantitative assessment of either over- or under-expression of a biomarker gene in a population of cells in a test sample relative to a standard (for example, a control sample). Such methods utilize reagents such as biomarker polynucleotides and biomarker antibodies.

10 [0074] In particular, the presence, absence or level of expression of a biomarker gene may be determined by measuring the amount of biomarker messenger RNA (mRNA), for example, by DNA-DNA hybridization, RNA-DNA hybridization, reverse transcription-polymerase chain reaction (PCR), real time quantitative PCR, differential display or TaqMan PCR; followed by comparing the results to a
15 reference based on a control sample (for example, samples from clinically-characterized patients and/or cell lines of a known genotype/phenotype). In one embodiment, microRNA expression or turnover may be measured. Hybridization, mass spectroscopy (e.g., MALDI-TOF or SELDI-TOF mass spectroscopy), serial analysis of gene expression or massive parallel signature sequencing assays can
20 also be performed. Non-limiting examples of hybridization assays include a singleplex or a multiplexed aptamer assay, a dot blot, a slot blot, an RNase protection assay, microarray hybridization, Southern or Northern hybridization analysis and in situ hybridization (e.g., Fluorescent *in situ* hybridization).

[0075] For example, these techniques find application in microarray-based assays
25 that can be used to detect and quantify the amount of biomarker gene transcript using cDNA- or oligonucleotide-based arrays. Microarray technology allows multiple biomarker gene transcripts and/or samples from different subjects to be analyzed in one reaction. Typically, mRNA isolated from a sample is converted into labeled nucleic acids by reverse transcription and optionally in vitro
30 transcription (cDNAs or cRNAs labelled with, for example, Cy3 or Cy5 dyes) and hybridized in parallel to probes present on an array. See, for example, Schulze et al, *Nature Cell Biol.*, 3:E190 (2001); and Klein et al., *J Exp Med*, 194:1625-1638

(2001), which are incorporated herein by reference in their entirety. Standard Northern analyses can be performed if a sufficient quantity of the test cells can be obtained. Utilizing such techniques, quantitative as well as size-related differences between biomarker transcripts can also be detected. In some embodiments, the expression level of the genes listed in Table 1 in the test sample are detected by using a human genome-wide array, a human lung tissue array or a custom array comprising polynucleotides of a plurality of genes in Table 1.

[0076] In some embodiments biomarkers are detected using reagents that specifically detect the biomarker. Such reagents may bind to a target gene or a target gene product (e.g., mRNA or protein), such that levels of the gene product may be quantified. Such reagents may be nucleic acid molecules that hybridize to the mRNA or cDNA of target gene products. Alternatively, the reagents may be molecules that label mRNA or cDNA for later detection, e.g., by binding to an array. The reagents may bind to proteins encoded by the genes of interest. For example, the reagent may be an antibody or a binding protein that specifically binds to a protein encoded by a target gene of interest. Alternatively, the reagent may label proteins for later detection, e.g., by binding to an antibody on a panel. In some embodiments, reagents are used in histology to detect histological and/or genetic changes in a sample.

[0077] The present invention provides isolated biomarker polynucleotides or variants thereof, which can be used, for example, as hybridization probes or primers ("biomarker probes" or "biomarker primers") to detect or amplify nucleic acids encoding a biomarker polypeptide, particularly a biomarker polypeptide encoded by a biomarker gene or polynucleotide selected from the group depicted in Table 1.

[0078] Nucleic acid molecules comprising nucleic acid sequences encoding the biomarker polypeptides or proteins of the invention, or genomic nucleic acid sequences from the biomarker genes (e.g., intron sequences, 5' and 3' untranslated sequences), or complements thereof (i.e., antisense polynucleotides), are collectively referred to as "biomarker genes," "biomarker polynucleotides" or "biomarker nucleic acid sequences" of the invention. The present invention also provides isolated biomarker polynucleotides or variants thereof, which can be

used, for example, as hybridization probes or primers (“biomarker probes” or “biomarker primers”) to detect or amplify nucleic acids encoding a biomarker polypeptide of the invention. The term “biomarker gene product” thus encompasses both mRNA as well as translated polypeptide as a gene product of a biomarker.

5 [0079] The isolated biomarker polynucleotide according to the invention may comprise flanking sequences (*i.e.*, sequences located at the 5’ or 3’ ends of the nucleic acid), which naturally flank the nucleic acid sequence in the genomic DNA of the organism from which the nucleic acid is derived. For example, in various
10 embodiments, the isolated biomarker polynucleotide can comprise less than about 5 kb, 4 kb, 3 kb, 2 kb, 1 kb, 0.5 kb or 0.1 kb of nucleotide sequences which naturally flank the coding sequence in genomic DNA of the cell from which the nucleic acid is derived. In other embodiments, the isolated biomarker polynucleotide is about 10-20, 21-50, 51-100, 101-200, 201-400, 401-750,
15 751-1000, or 1001-1500 bases in length.

[0080] In various embodiments, the biomarker polynucleotides of the invention are used as molecular probes in hybridization reactions or as molecular primers in nucleic acid extension reactions as described herein. In these instances, the biomarker polynucleotides may be referred to as biomarker probes and biomarker
20 primers, respectively, and the biomarker polynucleotides present in a sample which are to be detected and/or quantified are referred to as target biomarker polynucleotides. Two biomarker primers are commonly used in DNA amplification reactions and they are referred to as biomarker forward primer and biomarker reverse primer depending on their 5’ to 3’ orientation relative to the
25 direction of transcription.

[0081] In one embodiment, the invention encompasses methods of detecting genetic change in a biomarker gene (e.g., a mutation or a change in copy number). In another embodiment, the invention encompasses methods of detecting a change in the methylation of a biomarker gene.

30 [0082] A biomarker probe or a biomarker primer is typically an oligonucleotide which binds through complementary base pairing to a subsequence of a target biomarker polynucleotide. The biomarker probe may be, for example, a DNA

fragment prepared by amplification methods such as by PCR or it may be chemically synthesized. A double-stranded fragment may then be obtained, if desired, by annealing the chemically synthesized single strands together under appropriate conditions or by synthesizing the complementary strand using DNA
5 polymerase with an appropriate primer. Where a specific nucleic acid sequence is given, it is understood that the complementary strand is also identified and included as the complementary strand will work equally well in situations where the target is a double stranded nucleic acid. A nucleic acid probe is complementary to a target nucleic acid when it will anneal only to a single desired
10 position on that target nucleic acid under proper annealing conditions which depend, for example, upon a probe's length, base composition, and the number of mismatches and their position on the probe, and must often be determined empirically. Such conditions can be determined by those of skill in the art.

[0083] In one aspect of the invention, biomarkers may be detected in the test
15 sample or the control sample by gene expression profiling. In these methods, mRNA is prepared from a sample and mRNA expression levels are measured by reverse transcription quantitative polymerase chain reaction (RT-PCR followed with qPCR). RT-PCR is used to create a cDNA from the corresponding mRNA. The cDNA may be used in a qPCR assay to produce fluorescence as the DNA
20 amplification process progresses. By comparison to a standard curve, qPCR can produce an absolute measurement such as number of copies of mRNA per cell. Northern blots, microarrays, Invader assays, and RT-PCR combined with capillary electrophoresis may be used to measure expression levels of mRNA in a sample. Further details are provided, for example, in "Gene Expression Profiling: Methods
25 and Protocols," Richard A. Shimkets, editor, Humana Press, 2004 and US patent application 2010/0070191.

[0084] The invention encompasses an array comprising polynucleotides that hybridize to genes listed in Table 1. The array may comprise polynucleotides that hybridize to at least 2, at least 3, at least 4, at least 5, at least 10, at least 15, at least
30 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, least about 85, at least 87 or all genes listed in Table 1. In one embodiment, the polynucleotides are

immobilized on a solid surface. Examples of solid surfaces include paper, filter, nylon or other type of membrane, slide including glass slide, and chip (e.g., silicon, microarray chip). The polynucleotides may be single-stranded nucleic acid molecules (e.g., antisense oligonucleotides or fragments of cDNA). In some
5 embodiments, the array is not a human genome-wide array. Examples of human genome-wide array include, but are not limited to, Exon 1.0 ST, Gene 1.0 ST, U 95, U133, U133A 2.0, and U133 Plus 2.

[0085] In another aspect of the invention, detection of the biomarkers described herein may be accomplished by an immunoassay procedure. The immunoassay
10 typically includes contacting a test sample with an antibody that specifically binds to or otherwise recognizes a biomarker, and detecting the presence of the antibody/biomarker complex in the sample. The immunoassay procedure may be selected from a wide variety of immunoassay procedures known to those skilled in the art such as, for example, competitive or non-competitive enzyme-based
15 immunoassays, immunoprecipitation, enzyme-linked immunosorbent assays (ELISA), radioimmunoassay (RIA), immunofluorescence, immunohistochemistry (IHC), cytological assays and Western blots. Further, multiplex assays may be used, including antibody panels or arrays, wherein several desired antibodies are placed on a support, such as a glass bead or plate, and reacted or otherwise
20 contacted with the test sample or the control sample.

[0086] Antibodies used in these assays may be monoclonal or polyclonal, and may be of any type such as IgG, IgM, IgA, IgD and IgE. Monoclonal antibodies may be used to bind to a specific epitope offered by the biomarker molecule, and therefore may provide a more specific and accurate result. Antibodies may be
25 produced by immunizing animals such as rats, mice, rabbits and goats. The antigen used for immunization may be isolated from the samples or synthesized by recombinant protein technology. Methods of producing antibodies and of performing antibody-based assays are well-known to the skilled artisan and are described, for example, more thoroughly in *Antibodies: A Laboratory Manual*
30 (1988) by Harlow & Lane; *Immunoassays: A Practical Approach*, Oxford University Press, Gosling, J. P. (ed.) (2001) and/or *Current Protocols in Molecular Biology* (Ausubel et al.) which is regularly and periodically updated.

[0087] In certain embodiments, the present invention also provides “biomarker antibodies” including polyclonal, monoclonal, or recombinant antibodies, and fragments and variants thereof, that immunospecifically bind the respective biomarker proteins or polypeptides encoded by the genes or cDNAs (including polypeptides encoded by mRNA splice variants) as listed in Table 1.

[0088] Various chemical or biochemical derivatives of the antibodies or antibody fragments of the present invention can be produced using known methods. One type of derivative which is diagnostically useful as an immunoconjugate comprising an antibody molecule, or an antigen-binding fragment thereof, to which is conjugated a detectable label. However, in many embodiments, the biomarker antibody is not labeled but in the course of an assay, it becomes indirectly labeled by binding to or being bound by another molecule that is labeled. The invention encompasses molecular complexes comprising a biomarker antibody and a label, as well as immunocomplexes comprising a biomarker polypeptide, a biomarker antibody, and immunocomplexes comprising a biomarker polypeptide, a biomarker antibody, and a label.

[0089] Examples of detectable substances or detectable labels include various enzymes, prosthetic groups, fluorescent materials, luminescent materials, bioluminescent materials, and radioactive materials. Examples of suitable enzymes include horseradish peroxidase, alkaline phosphatase, beta-galactosidase and acetylcholinesterase. Examples of suitable prosthetic group complexes include streptavidin/biotin and avidin/biotin. Examples of suitable fluorescent materials include umbelliferones, fluoresceins, fluorescein isothiocyanate, rhodamines, dichlorotriazinylamine fluorescein, dansyl chloride, phycoerythrins, Alexa Fluor 647, Alexa Fluor 680, DilC₁₉(3), Rhodamine Red-X, Alexa Fluor 660, Alexa Fluor 546, Texas Red, YOYO-1 + DNA, tetramethylrhodamine, Alexa Fluor 594, BODIPY FL, Alexa Fluor 488, Fluorescein, BODIPY TR, BODIPY TMR, carboxy SNARF-1, FM 1-43, Fura-2, Indo-1, Cascade Blue, NBD, DAPI, Alexa Fluor 350, aminomethylcoumarin, Lucifer yellow, Propidium iodide, or dansylamide. An example of a luminescent material is luminol. Examples of bioluminescent materials include green fluorescent proteins, modified green

fluorescent proteins, luciferase, luciferin, and aequorin. Examples of suitable radioactive material include ^{125}I , ^{131}I , ^{35}S or ^3H .

[0090] Immunoassays for biomarker polypeptides will typically comprise incubating a sample, such as a biological fluid, a tissue extract, freshly harvested
5 cells, or lysates of cells, in the presence of a detectably labeled antibody capable of identifying biomarker gene products or conserved variants or peptide fragments thereof, and detecting the bound antibody by any of a number of techniques well-known in the art. One way of measuring the level of biomarker polypeptide with a specific biomarker antibody of the present invention is by enzyme immunoassay
10 (EIA) such as an enzyme-linked immunosorbent assay (ELISA) (Voller, A. *et al.*, *J. Clin. Pathol.* 31:507-520 (1978); Butler, J.E., *Meth. Enzymol.* 73:482-523 (1981); Maggio, E. (ed.), *Enzyme Immunoassay*, CRC Press, Boca Raton, FL, 1980). The enzyme, either conjugated to the antibody or to a binding partner for the antibody, when later exposed to an appropriate substrate, will react with the
15 substrate in such a manner as to produce a chemical moiety which can be detected, for example, by spectrophotometric, or fluorimetric means.

[0091] The biological sample may be brought in contact with and immobilized onto a solid phase support or carrier such as nitrocellulose, or other solid support which is capable of immobilizing cells, cell particles or soluble proteins. The
20 support may then be washed with suitable buffers followed by treatment with the detectably labeled biomarker antibody. The solid phase support may then be washed with the buffer a second time to remove unbound antibody. The amount of bound label on solid support may then be detected by conventional means. A well known example of such a technique is Western blotting.

25 [0092] In various embodiments, the present invention provides compositions comprising labelled biomarker polynucleotides, or labelled biomarker antibodies to the biomarker proteins or polypeptides, or labeled biomarker polynucleotides and labeled biomarker antibodies to the biomarker proteins or polypeptides according to the invention as described herein.

30 [0093] Antibodies and other reagents may also be used to detect post-translational modifications (e.g., methylation, acetylation, farnesylation, biotinylation, stearoylation, formylation, myristoylation, palmitoylation,

geranylgeranylation, pegylation, phosphorylation, sulphation, glycosylation, sugar modification, lipidation, lipid modification, ubiquitination, sumolation, disulphide bonding, cysteinylolation, oxidation, glutathionylation, carboxylation, glucuronidation, and deamidation) of biomarker proteins or biomarker polypeptides.

5 [0094] The invention encompasses a panel comprising antibodies that bind to proteins encoded by genes listed in Table 1. The panel may comprise antibodies that bind to proteins encoded by at least 2, at least 3, at least 4, at least 5, at least 10, at least 15, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45,
10 at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, least about 85, at least 87 or all genes listed in Table 1. In one embodiment, the panel of antibodies is immobilized on a solid surface. Examples of solid surfaces include microspheres, plates, wells, slides, and beads (e.g., protein A or protein G agarose).

15 [0095] In addition to antibody-based techniques, the biomarkers described herein may also be detected and quantified by mass spectrometry. Mass spectrometry is a method that employs a mass spectrometer to detect ionized protein markers or ionized peptides as digested from the protein markers by measuring the mass-to-charge ratio (m/z). Labelling of biomarkers (along with other proteins) with stable
20 heavy isotopes (deuterium, carbon-13, nitrogen-15, and oxygen-18) can be used in quantitative proteomics. These are either incorporated metabolically in sample cells cultured briefly in vitro, or directly in samples by chemical or enzymatic reactions. Light and heavy labelled biomarker peptide ions segregate and their intensity values are used for quantification. For example, analytes may be
25 introduced into an inlet system of the mass spectrometer and ionized in an ionization source, such as a laser, fast atom bombardment, plasma or other suitable ionization sources known to the art. The generated ions are typically collected by an ion optic assembly and introduced into mass analyzers for mass separation before their masses are measured by a detector. The detector then translates
30 information obtained from the detected ions into mass-to-charge ratios.

[0096] The invention also encompasses methods that involve measuring the activity of a biomarker (e.g., enzymatic activity). Examples of enzymatic activity

include, without limitation, kinase, phosphatase, protease, ubiquitination, oxidase and reductase activity.

[0097] The invention also provides compositions comprising biomarker polynucleotides, biomarker polypeptides, or biomarker antibodies according to the invention as described herein in the various embodiments. The invention further provides diagnostic or detection reagents for use in the methods of the invention, for example, reagents for flow cytometry and/or immunoassays that comprise fluorochrome-labeled antibodies that bind to one of the biomarker polypeptides of the invention.

[0098] In one embodiment, the invention provides diagnostic or detection reagents that comprise one or more biomarker probes, or one or more biomarker primers. A diagnostic reagent may comprise biomarker probes and/or biomarker primers from the same biomarker gene or from multiple biomarker genes. In another embodiment, the invention also provides diagnostic compositions that comprise one or more biomarker probes and target biomarker polynucleotides, or one or more biomarker primers and target polynucleotides, or biomarker primers, biomarker probes and biomarker target polynucleotides. In some embodiments, the diagnostic compositions comprise biomarker probes and/or biomarker primers and a sample suspected to comprise biomarker target polynucleotides. Such diagnostic compositions comprise biomarker probes and/or biomarker primers and the nucleic acid molecules (including RNA, mRNA, cRNA, cDNA, and/or genomic DNA) of a subject in need of a diagnosis/prognosis of lung cancer.

In silico Analysis and Computer Readable Media

[0099] Biomarkers and gene signatures of the invention may be predicted based on gene expression patterns in lung cancer, including stages 1 and 2 adenocarcinoma and stages 1 and 2 squamous cell carcinoma. In some embodiments, biomarker and gene signature prediction comprises gene expression patterns in control (e.g., non-tumor) biological samples. A heterogeneous ensemble learning approach may be used to classify samples based on their gene expression profiles. Such an approach may combine predictions from different approaches that use genes, gene set-derived features and/or causal network-derived features in order to get a classification and a prediction confidence for each

classified sample. Methods that may be used to generate biomarkers and gene signatures of the invention include shrunken centroids, factor rotation, logistic regression models, network-based approaches, disease module-based approaches, linkage methods, modularity or pathway-based methods and diffusion-based methods.

[0100] The biological data (such as training data and test data) used in these methods may be drawn from the literature, databases (including data from preclinical, clinical and post-clinical trials of pharmaceutical products or medical devices), genome databases (genomic sequences and expression data, *e.g.*, Gene Expression Omnibus by National Center for Biotechnology Information or ArrayExpress by European Bioinformatics Institute (Parkinson et al. 2010, Nucl. Acids Res., doi: 10.1093/nar/gkq1040. Pubmed ID 21071405)), commercially available databases (*e.g.*, Gene Logic, Gaithersburg, MD, USA) or experimental work. In one embodiment, the REACTOME, KEGG or BIOCARTA pathway gene set collections from the Broad Institute (Cambridge, MA) may be used. The data may be related to nucleic acid (*e.g.*, absolute or relative quantities of specific DNA or RNA species, changes in DNA sequence, RNA sequence, changes in tertiary structure, or methylation pattern as determined by sequencing, hybridization - particularly to nucleic acids on microarray, quantitative polymerase chain reaction, or other techniques known in the art), protein/peptide (*e.g.*, absolute or relative quantities of protein, specific fragments of a protein, peptides, changes in secondary or tertiary structure, or posttranslational modifications as determined by methods known in the art) and functional activities (*e.g.*, enzymatic activities, proteolytic activities, transcriptional regulatory activities, transport activities, binding affinities to certain binding partners) under certain conditions, among others. Modifications, including posttranslational modifications of protein or peptide, can include, but are not limited to, methylation, acetylation, farnesylation, biotinylation, stearoylation, formylation, myristoylation, palmitoylation, geranylgeranylation, pegylation, phosphorylation, sulphation, glycosylation, sugar modification, lipidation, lipid modification, ubiquitination, sumolation, disulphide bonding, cysteinylolation, oxidation, glutathionylation, carboxylation, glucuronidation, and deamidation. In addition, a protein can be modified

posttranslationally by a series of reactions such as Amadori reactions, Schiff base reactions, and Maillard reactions resulting in glycated protein products.

[0101] The test data sets may be processed and have their quality controlled separately, together if they are obtained from the same technology platform (e.g.,
5 an Affymetrix platform). For example, raw data files may be read by the ReadAffy function of the affy package (Gautier et al., *Bioinformatics*, 20:307-315 (2004)) belonging to Bioconductor (Gentleman et al., *Genome Biol*, 5(10):R80 (2004)) in R (R Development Core Team, *R: A Language and Environment for Statistical Computing*. 2007). The quality may be controlled by:

- 10 1. Generating RNA degradation plots (using the AffyRNAdeg function of the affy package (Gautier, 2004)), NUSE and RLE plots (using the function affyPLM) (Brettschneider et al., *Technometrics*, 50(3):241-264 (2008)), calculating the MA(RLE) values;
2. Excluding arrays from the training datasets that fall below a set of
15 thresholds on the quality control checks or that do not correspond to the test parameters;
or both.

Arrays passing quality control checks may be normalized using the germa algorithm (Wu et al., *Journal of the American Statistical Association*, 99:909
20 (2004)). If the datasets were obtained from a database, the samples classification may be obtained from the series matrix file of the same database for each dataset. The output of this part of the method may consist of: a gene expression matrix on training samples and test samples, probesets, and the class information for the training samples.

25 [0102] Non-limiting examples of methods that may be used to generate predictions are: transformation invariant (TranInv) (U.S. Provisional Patent Application entitled "Systems and Methods for Generating Biomarker Signatures with Integrated Bias Correction and Class Prediction," filed concurrently with the instant application and having the attorney docket no. 106500-0032-001), dual
30 ensemble (Yang et al., *Current Bioinformatics*, 5(4):296-308 (2010)), generalized simulated annealing (Tsallis and Stariolo, *Physica A: Statistical Mechanics and Its Applications*, 233(1):395-406 (1996); Xiang and Gong, *Physical Review E*,

62(3):4473 (2000); Xiang et al., *Physics Letters A*, 233(3):216-220 (1997); Xiang et al., *The Journal of Physical Chemistry A*, 104(12):2746-2751 (2000)), T-filter, CORG (Chuang et al., *Mol Syst Biol*, 3:140 (2007)), single and pairs, dual bagging, forward learning, NPA (network perturbation amplitude) (see, e.g., International Patent Application No. PCT/EP2012/061035, filed June 11, 2012 and U.S. Provisional Patent Application entitled "Systems and Methods Relating to Network-Based Biomarker Signatures," filed concurrently with the instant application and having the attorney docket no. 106500-0022-001) and Laplacian based learning. Each of the foregoing patent applications and publications are incorporated herein by reference in their entirety.

[0103] Generalized simulated annealing may be modified for binary functions. In one embodiment, a dual binary generalized simulated annealing based method may be used (DualGensemble) (U.S. Provisional Patent Application entitled "Systems and Methods for Generating Biomarker Signatures with Integrated Dual Ensemble and Simulated Annealing Techniques," filed concurrently with the instant application, incorporated herein by reference in its entirety and having the attorney docket no. 106500-0031-001). T-filter is a method of filtering genes based on the t-test by setting P-value and fold-change thresholds. CORG may be modified by calculating activity scores by leveraging the F-test instead of the T-test. CORG may also be combined with SVM. Dual bagging is a combination of bagging (Breiman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed., ed. T. Hastie, R. Tibshirani, and J. Friedman, (2009)) and the random subspace method (Bryll, *Pattern Recognition*, 20(6):1291-1302 (2003); Ho, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832-844 (1998); Skurichina, *Pattern Analysis and Applications*, 5(2):121-135 (2002)).

[0104] The single and pairs method may include the following steps:

1. Select a threshold a_0 .
2. For each pair of genes compute the leave-one-out cross-validation of a quadratic discriminant analysis and record the accuracy. Save the pair of genes if this accuracy is less than a_0 .

3. For each gene, compute the a_0 and $1 - a_0$ quantile of the values in each class to discriminate the classes A and B, $Q_{a_0}(A)$, $Q_{1-a_0}(A)$, $Q_{a_0}(B)$, $Q_{1-a_0}(B)$. Select the gene if either: $Q_{a_0}(A) > Q_{1-a_0}(B)$ or $Q_{1-a_0}(A) < Q_{a_0}(B)$.
4. Use the obtained list to train a classification algorithm on the reduced feature space.
5. Choose a_0 by cross-validation.

[0105] The forward learning method may include the following steps:

Set IN to the empty list, choose N (for example, 20, 100 or 200).

For $n = 1, \dots, N$ do

- a. For each gene, g, not in IN, compute a randomForest (ntree=500) on the subspace corresponding to $\{IN, g\}$, record the out-of-the bag true positive rates (TPr) and true negative rates (TNr) and compute the g-performance $\sqrt{TPr * TNr}$.
 - b. Select the gene, g_{max} , for which the g-performance is maximum and add it to the list $IN := \{IN, g_{max}\}$.
 - c. Then train a classification algorithm on the subspace given by IN.
- N is chosen by cross-validation.

[0106] The Laplacian based learning method may include the following steps:

1. Compute Spearman or Pearson correlation between the samples based on their gene expression profiles for both test and training data. Normalize the distance matrix obtained from the correlation matrix ($K_{ij} = K_{ij} / \sqrt{K_{ii} * K_{jj}}$).
2. Compute the k-nearest neighbors of each sample (k chosen by cross-validation, usually $k = 2, 3, 4, 5$).
3. Define a graph with samples as nodes and put an edge between neighbors.
4. Create the (combinatorial) Laplacian of the graph and get its generalized inverse G, which is a positive definite kernel.
5. Extract main kernel principal components (KPC) from G and train a Svm on it. The number of KPC's is chosen with rdetools package function rde (Braun et al., *The Journal of Machine Learning Research*, 9:1875-1908 (2008)).

6. Train a SVM with the training samples and get performance by cross-validation.

7. Predict the test cases.

5 [0107] Network-based analysis can be combined with machine learning methods to generate predictions, for example, combining any one of CORG, dual bagging or T-filter with a network-based analysis.

[0108] In some embodiments, methods used to generate predictions are further combined with another classification method (e.g., a method that is used for cross-validation). Non-limiting examples of classification methods include PAMR
10 (Tibshirani et al., *Proc Natl Acad Sci USA*, 99(10):6567-6572 (2002)), RandomForest (Breiman, *Machine Learning*, 45(1):5-32 (2001)), Linear Discrimination Analysis (LDA), Eigengene-based Linear Discrimination Analysis (ELDA), Principal Components Analysis (PCA), Recursive Partitioning Tree (RPART), Support Vector Machine (SVM), K-Nearest Neighbors (KNN) (Bishop,
15 Neural Networks for Pattern Recognition, ed. O.U. Press, 1995) and Partial Least Squares Discriminant Analysis (PLS.DA). In one embodiment, a network-based analysis that uses NPA may be combined with SVM (U.S. Provisional Patent Application entitled "Systems and Methods relating to Network-based Biomarker Signatures," filed concurrently with the instant application, incorporated herein by
20 reference in its entirety and having the attorney docket no. 106500-0022-001).

[0109] In one embodiment, these methods may further include a step of oversampling to balance classes. The methods may include a step of filtering genes based on a simple T-test between the categories to be classified. The filtering step may reduce the number of genes to less than 1,500 or less than 2,000.

25 [0110] After predictions are generated by several methods, a vote may be made to obtain the classification as well as the confidence for the prediction of each sample of the sample set. If a method provides cross-validation results far below the other methods, it may be excluded. Such additional steps are contemplated in the methods of the invention.

30 [0111] The union of the gene signatures extracted by these methods may be considered as the larger gene signature. A weight may be given to genes to take into consideration the number of times they appear in a list. See, for example, in

Table 1 the column “present/total lists” which shows the number of times each gene appears in one of the predicted gene signatures.

[0112] The genes obtained by these methods may be mapped to gene symbols using any suitable platform, for example, the *Confero* platform (Hermida et al.,
5 *Confero: an Integrated Contrast and Gene Set Platform for Computational Analysis and Biological Interpretation of Omics Data*. submitted, 2012).

[0113] The numerical methods for generating the gene signatures of the invention may include a testing step and confidence statistics for the genes. The testing step (or phase) is an exemplary use of the gene signature in carrying out the
10 claimed method.

[0114] The invention encompasses a method for classifying a test sample as stage 1 lung adenocarcinoma, stage 2 lung adenocarcinoma, stage 1 squamous cell carcinoma or stage 2 squamous cell carcinoma, the method comprising: measuring the expression levels of at least 2 genes listed in Table 1 in a test sample; and
15 applying one or more network-based methods, one or more machine-learning based methods, or a combination of the foregoing methods to the expression levels to obtain a classification of the test sample as stage 1 lung adenocarcinoma, stage 2 lung adenocarcinoma, stage 1 squamous cell carcinoma or stage 2 squamous cell carcinoma. In some embodiments, the expression levels of at least 2, at least 3, at
20 least 4, at least 5, at least 10, at least 15, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, least about 85, at least 87 or all genes listed in Table 1 are measured. In some embodiments, the classifier has been trained by *in silico* analysis or one or more feature selection and classification algorithms.

[0115] One aspect of the invention encompasses a list of one or more biomarkers or gene signatures of the invention stored on a computer readable medium. The absence, presence, activity or expression level of a biomarker in a biological sample (such as a control sample or a test sample) may also be stored on the computer readable medium. The computer readable medium may also include
30 information that identifies the sample. The computer readable medium may also include a computer program product.

[0116] The computer program product may include a classifier based on at least two genes listed in Table 1. The classifier may be based on at least 2, at least 3, at least 4, at least 5, at least 10, at least 15, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, least about 85, at least 87 or all genes listed in Table 1.

[0117] Optionally, the classifier is trained by *in silico analysis* or one or more feature selection and classification algorithms. In some embodiments, the classifier is trained by one or more algorithms selected from the group consisting of dual ensemble, generalized simulated annealing, T-filter, CORG, CORG combined with support vector machine, dual bagging, single and pairs, forward learning, Laplacian based learning and learning method based on network perturbation amplitude. The classifier may be trained with at least the data in Gene Expression Omnibus datasets GSE2109, GSE10245, GSE18842 and GSE37745.

Devices and Kits

[0118] One aspect of the invention encompasses devices useful for performing methods of the invention. For example, the devices may be used for diagnosing, classifying and/or grading lung cancer. The devices can comprise means for detecting the expression level of at least 2 of the genes listed in Table 1 or the level of at least 2 gene products of such genes in a test sample. Such means may include components for performing one or more methods of nucleic acid extraction, nucleic acid amplification, nucleic acid detection, protein isolation and/or protein detection. Such components may include one or more of an amplification chamber (for example a thermal cycler), a plate reader, robotic sample handling components, a capillary electrophoresis apparatus, a spectrophotometer, a mass spectrometer and/or a chip reader. These components can obtain data that reflects the expression level of the genes being analyzed. In some embodiments, the devices can comprise means for detecting at least 2, at least 3, at least 4, at least 5, at least 10, at least 15, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, at least 84 or all genes listed in Table 1. In some embodiments, the devices can comprise means for detecting the gene products of at least 2, at least 3, at least 4, at least 5, at least 10, at least 15, at least 20, at least 25, at least 30, at

least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, at least 84 or all genes listed in Table 1.

[0119] The devices optionally comprise a means for identifying a given test sample, and of linking the results obtained to that sample. Such means can include manual labels, barcodes, and other indicators which can be linked to a sample container or receptacle. Identification means may optionally be included in the sample itself, for example where an encoded particle is added to the sample. The results may be linked to the sample, for example in a computer memory that contains a sample designation and a record of expression levels obtained from the sample. Linkage of the results to the sample can also include a linkage to a particular sample container or receptacle in the device, which is also linked to the sample identity.

[0120] The devices may comprise an excitation and/or a detection means. Any instrument that provides a wavelength that can activate a label (e.g., fluorophore, fluorochrome and fluorescent dye) used on a detection reagent and is shorter than the emission wavelength(s) to be detected can be used for excitation. Examples of excitation sources include a broadband ultraviolet light source such as a deuterium lamp with an appropriate filter, the output of a white light source such as a xenon lamp or a deuterium lamp after passing through a monochromator to extract out the desired wavelength(s), a continuous wave (cw) gas laser, a solid state diode laser, or any pulsed lasers. Emitted light can be detected through any suitable component or technique; many suitable approaches are known in the art. For example, a fluorimeter or spectrophotometer may be used to detect whether the test sample emits light of a wavelength characteristic of a label used in a method of the invention.

[0121] The devices may comprise a means for correlating the expression levels of the genes being analyzed with an lung cancer status, prognosis, grade and/or classification. Such means may comprise one or more of a variety of correlative techniques, including lookup tables, algorithms, multivariate models, and linear or nonlinear combinations of expression models or algorithms, such as any of the *in silico* and machine learning methods described above. The expression levels may be converted to one or more biomarker scores, indicating that the individual

providing the sample is not suffering from adenocarcinoma or squamous cell carcinoma or is suffering from stage 1 adenocarcinoma, stage 2 adenocarcinoma, stage 1 squamous cell carcinoma or stage 2 squamous cell carcinoma. The models and/or algorithms can be provided in computer readable format.

5 [0122] The devices may also comprise output means for outputting the lung cancer status, prognosis, grade and/or classification. Such output means can take any form which transmits the results to an individual and/or a healthcare provider, and may include a monitor, a display, and/or a printer. Output means may record the results to a computer readable medium. The device may use a computer
10 system for performing one or more of the steps provided.

[0123] In one embodiment, a device of the invention comprises means for detecting the expression level of at least 2 genes listed in Table 1 in a test sample; means for correlating the expression level with a classification of the lung cancer as stage 1 adenocarcinoma, stage 2 adenocarcinoma, stage 1 squamous cell
15 carcinoma or stage 2 squamous cell carcinoma; and means for outputting the lung cancer classification. In some embodiments, the device comprises means for detecting the expression level of at least 2, at least 3, at least 4, at least 5, at least 10, at least 15, at least 20, at least 25, at least 30, at least 35, at least 40, at least 45, at least 50, at least 55, at least 60, at least 65, at least 70, at least 75, at least 80, at
20 least 84 genes or all genes listed in Table 1.

[0124] Another aspect of the invention encompasses kits for practicing the methods of the invention. Such kits may be used for classifying and grading lung cancer or for assessing the prognosis of lung cancer in an individual. The kits can be used for clinical diagnosis and/or laboratory research. In one embodiment, a kit
25 comprises in one or more containers one or more reagents that detect expression levels of genes that serve as biomarkers of lung cancer in a test sample. Preferably, the kit also comprises instructions in any tangible medium (e.g., written, tape, CD-ROM, DVD) on the use of the detection reagent(s) in one or more methods of the invention.

30 [0125] For nucleic acid-based methods (for example, amplification assays, hybridization assays, sequencing or polymerase chain reactions), a detection reagent in the kit may comprise at least one polynucleotide, probe, and/or primer

specific for the stage 1 adenocarcinoma, stage 2 adenocarcinoma, stage 1 squamous cell carcinoma and/or stage 2 squamous cell carcinoma the genes listed in Table 1. The nucleic-acid based detection reagents may comprise sequences complementary to a portion of the signature genes or sequences that are portions of the signature genes. Such a kit may optionally provide in separate containers enzymes and/or buffers for reverse transcription, *in vitro* transcription, and/or DNA polymerization, nucleotides, and/or labeled nucleotides.

[0126] For protein-based methods, such as immunoassays, a detection reagent in the kit may comprise a biomarker antibody, which may be labeled or labelable.

10 The antibodies may bind to proteins encoded by stage 1 adenocarcinoma, stage 2 adenocarcinoma, stage 1 squamous cell carcinoma and/or stage 2 squamous cell carcinoma the genes listed in Table 1. In one embodiment, the detection reagents recognize a post-translational modification (e.g., methylation, acetylation, farnesylation, biotinylation, stearylation, formylation, myristoylation, 15 palmitoylation, geranylgeranylation, pegylation, phosphorylation, sulphation, glycosylation, sugar modification, lipidation, lipid modification, ubiquitination, sumolation, disulphide bonding, cysteinylolation, oxidation, glutathionylation, carboxylation, glucuronidation, and deamidation) of a protein encoded by a gene selected from the genes listed in Table 1. For protein-based methods that involve 20 measuring the activity of a biomarker (e.g., enzymatic activity), the kit may include a substrate for the biomarker and a detection reagent that recognizes the products and/or byproducts of the activity being measured. Such a kit may optionally provide, in separate containers, buffers, secondary antibodies, signal generating accessory molecules, and/or labeled secondary antibodies, including 25 fluorochrome-labeled secondary antibodies. The kit may also include unlabeled or labeled antibodies to various cell surface antigens which can used for identification or sorting of subpopulations of cells.

[0127] The detection reagents may be labeled or labelable by one or more detectable labels. Examples of detectable labels include, without limitation, 30 radiolabels (e.g. radioactive nuclides), dyes, fluorescent proteins or materials (e.g., fluorochromes, fluorophores, fluorescein and rhodamine), luminescent proteins or materials, bioluminescent proteins or materials (e.g., luciferase, aequorin and

luciferin), enzymes (e.g., beta-galactosidase, alkaline phosphatase, horseradish peroxidase and acetylcholinesterase) and prosthetic groups (e.g., biotin, streptavidin and avidin).

5 [0128] The detection reagents in the kit may be immobilized on a solid surface or packaged separately with reagents to immobilize them on a solid surface.

[0129] Also included in the kit may be positive and negative controls for the methods of the invention. The positive and/or negative controls included in a kit can be nucleic acids, polypeptides, cell lysate, cell extract, whole cells from patients, or whole cells from cell lines.

10 **Example 1 Generation of Lung Cancer Gene Signatures**

[0130] This example is for the purpose of illustration only and is not to be construed as limiting the scope of the invention in any way. A heterogeneous ensemble learning approach aimed at classifying samples based on their gene expression profile is applied to extract genes whose expression levels allow
15 determination of adenocarcinoma from squamous cell carcinoma lung biopsies and determination of stage 1 from stage 2 for each type of carcinoma. In summary, predictions from different approaches that use genes and gene sets-derived features are combined to get the most accurate classifier possible. Gene lists are extracted from these methods and combined to generate the list presented in Table 1.

20 [0131] A schematic overview of the data and strategy used to generate the list given in Table 1 is given below:

1. Public datasets and a dataset downloaded from the SBV diagnostic signature challenge (<http://sbvimprover.com>) are used as the source of gene expression data.

25 The following public datasets are downloaded from the Gene Expression Omnibus (GEO) (<http://www.ncbi.nlm.nih.gov/geo/>) repository:

- GSE2109 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2109)
- GSE10245 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE10245)
- GSE18842 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18842)
- 30 •GSE37745 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE37745)

2. As both training datasets are on the same Affymetrix platform as the test dataset (HGU-133 + 2), they are processed and have their quality controlled together. In summary, raw data files are read by the ReadAffy function of the affy package (Gautier, 2004) belonging to Bioconductor (Gentleman, 2004) in R (R Development Core Team, 2007), and the quality is controlled by:
- a. generating RNA degradation plots (with the AffyRNAdeg function of the affy package), NUSE and RLE plots (with the function affyPLM (Brettschneider, 2008)), and calculating the MA(RLE) values;
 - b. excluding arrays from the training datasets that fell below a set of thresholds on the quality control checks or that are duplicated in the above datasets; and
 - c. normalizing arrays that pass quality control checks using the germa algorithm (Wu, 2004). Training set sample classifications are obtained from the series matrix file of the GEO database for each dataset.
- 15 Arrays GSM926687_PA025_203B_081029_HG-U133_Plus_2_.CEL,
 GSM926694_PA025_224A_081028_HG-U133_Plus_2_.CEL,
 GSM926700_PA025_260A_081029_HG-U133_Plus_2_.CEL,
 GSM926706_PA025_296A_081211_HG-U133_Plus_2_.CEL,
 GSM926712_PA025_344A_081106_HG-U133_Plus_2_.CEL,
 20 GSM926717_PA025_98A_081029_HG-U133_Plus_2_.CEL,
 GSM926718_PA025_human_112B_070912.CEL,
 GSM926721_PA025_human_131A_070228.CEL,
 GSM926728_PA025_human_175B_070228.CEL,
 GSM926747_PA025_human_295B_070912.CEL,
 25 GSM926758_PA025_human_331A_070914.CEL,
 GSM926761_PA025_human_342A_070907.CEL,
 GSM926762_PA025_human_347A_070907.CEL,
 GSM926770_PA025_human_37A_070228.CEL,
 GSM926776_PA025_human_88A_070228.CEL,
 30 GSM926787_PA117_130A_101126_HG-U133_Plus_2_.CEL,
 GSM926801_PA117_173B_101130_HG-U133_Plus_2_.CEL,
 GSM926826_PA117_285A_101210_HG-U133_Plus_2_.CEL,

GSM926840_PA117_336A_101202_HG-U133_Plus_2_.CEL,
GSM926843_PA117_353A_101208_HG-U133_Plus_2_.CEL,
GSM926854_PA117_58A_101123_HG-U133_Plus_2_.CEL,
GSM926856_PA117_5A_101118_HG-U133_Plus_2_.CEL,
5 GSM926866_PA117_7A_101118_HG-U133_Plus_2_.CEL,
GSM926872_PA117_96A_101126_HG-U133_Plus_2_.CEL,
GSM102451.CEL, GSM102455.CEL, GSM138003.CEL, GSM203732.CEL,
GSM231874.CEL, GSM467024.CEL, GSM76585.CEL, GSM76595.CEL,
GSM88997.CEL, GSM926766_PA025_human_365A_070907.CEL,
10 GSM926739_PA025_human_256A_070926.CEL,
GSM926738_PA025_human_255A_070926.CEL, GSM89060.CEL,
GSM76590.CEL, GSM76587.CEL, GSM53170.CEL, GSM53167.CEL,
GSM46817.CEL, GSM467030.CEL, GSM467029.CEL, GSM466975.CEL,
GSM466952.CEL, GSM38104.CEL, GSM38103.CEL, GSM203641.CEL,
15 GSM152757.CEL, GSM152681.CEL, GSM152670.CEL, GSM152624.CEL,
GSM138002.CEL, GSM137945.CEL, GSM137931.CEL, GSM137916.CEL,
GSM137910.CEL, GSM117763.CEL, GSM117632.CEL, GSM117610.CEL,
GSM102555.CEL, GSM102553.CEL, GSM102548.CEL, GSM102512.CEL,
GSM102507.CEL, GSM102447.CEL, GSM117763.CEL, GSM152757.CEL,
20 GSM258570.CEL, GSM258591.CEL, GSM258594.CEL, GSM258597.CEL,
GSM258601.CEL, GSM258603.CEL, GSM258606.CEL, GSM466973.CEL,
GSM466982.CEL, GSM467029.CEL, GSM46833.CEL, GSM53167.CEL,
GSM53170.CEL, GSM926683_PA025_19A_081028_HG-
U133_Plus_2_.CEL, GSM926692_PA025_217A_081029_HG-
25 U133_Plus_2_.CEL, GSM926703_PA025_282A_081029_HG-
U133_Plus_2_.CEL, GSM926707_PA025_301A_081029_HG-
U133_Plus_2_.CEL, GSM926708_PA025_302A_081106_HG-
U133_Plus_2_.CEL, GSM926733_PA025_human_207A_070912.CEL
,GSM926734_PA025_human_229A_070912.CEL,
30 GSM926735_PA025_human_234A_070912.CEL,
GSM926741_PA025_human_265A_071005.CEL,
GSM926745_PA025_human_284A_070912.CEL,

GSM926750_PA025_human_300A_070919.CEL,
 GSM926755_PA025_human_317A_070926.CEL,
 GSM926760_PA025_human_33A_070509.CEL,
 GSM926769_PA025_human_373B_070907.CEL,
 5 GSM926772_PA025_human_48A_070509.CEL,
 GSM926784_PA117_116A_101126_HG-U133_Plus_2_.CEL,
 GSM926786_PA117_128A_101126_HG-U133_Plus_2_.CEL,
 GSM926798_PA117_164A_101210_HG-U133_Plus_2_.CEL,
 GSM926803_PA117_179A_101130_HG-U133_Plus_2_.CEL,
 10 GSM926809_PA117_190B_101202_HG-U133_Plus_2_.CEL,
 GSM926818_PA117_241B_101203_HG-U133_Plus_2_.CEL,
 GSM926836_PA117_321A_101203_HG-U133_Plus_2_.CEL,
 GSM926861_PA117_74A_101215_HG-U133_Plus_2_.CEL,
 GSM926862_PA117_75A_101126_HG-U133_Plus_2_.CEL,
 15 GSM102512.CEL, GSM138002.CEL, GSM467032.CEL,
 GSM926693_PA025_223A_081028_HG-U133_Plus_2_.CEL,
 GSM926756_PA025_human_319A_070914.CEL,
 GSM926849_PA117_41A_101123_HG-U133_Plus_2_.CEL,
 GSM926851_PA117_52A_101123_HG-U133_Plus_2_.CEL,
 20 GSM102548.CEL, GSM102553.CEL, GSM117610.CEL, GSM38103.CEL,
 GSM46868.CEL, GSM46936.CEL, GSM46941.CEL, GSM76587.CEL,
 GSM76590.CEL, GSM88962.CEL were not used for further analysis.

The output at this point comprises a gene expression matrix X on 410 samples
 (260 training samples and 150 test samples) and 54675 probesets, and the class
 25 information for the training samples.

3. Features selection and classification algorithm(s) used for prediction of a
 gene signature follow the illustrated strategy of Figure 1:

Briefly, a set of feature selection and classification algorithms are used to obtain a
 number of classifications for each test sample. Each method has defined input and
 30 output:

- INPUT: gene expression matrix $X_{n \times p}$ on n samples and p genes, training
 samples and test samples, and the class information for the training samples

- OUTPUT: Class prediction for each test sample and a list of genes involved.

[0132] Prior to applying features selection and classification methods, the following steps are performed: (1) oversampling is, optionally, used to balance classes in the training dataset; (2) mapping probe sets to gene symbols (Entrez gene ids) using *Confero* platform (Hermida, 2012); and (3) optionally filtering the genes in the matrix based on simple T-test between the categories to be classified so that less than 1500 genes (for Dual Ensemble or T-filter methods) or less than 2000 genes (for the other methods) remain.

10 [0133] Cross validation of the results is performed using any of the following supervised methods:

- PAMR (Tibshirani, 2002)
- RandomForest (Breiman, 2001)
- Linear Discriminant Analysis (LDA)
- 15 • Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN) (Bishop, 1995)
- Partial Least Squares Discriminant Analysis (PLS.DA)

[0134] The following methods are used to generate predictions:

a. Dual Ensemble method.

20 This dual ensemble method builds ensemble of multiple classification algorithms applied in randomly perturbed data. The diversity of the ensemble classifier is imposed by using different classification algorithms and is further enhanced by data-level perturbation. See, e.g., Yang, 2010. A molecular profile of a training dataset, T0.train and its associated phenotype cl.train (control and treatment) are used as input. The molecular profile of the test set T0.test is used to predict the phenotype cl.test.

b. T-filter method

25 Genes are filtered based on t-test to obtain a list of N genes, by setting P-value and fold-change thresholds. Thresholds are decreased (resp. increase) automatically if the list size is over N. Any M is trained on the resulting subspace. N is chosen by cross-validation.

30 **c. CORG-modified method**

This method is modified from CORG method (Chuang, 2007) as activity scores are calculated by leveraging F-test instead of T-test. It uses the c2.cp gene sets collection from the Broad Institute (Cambridge, MA) (Reactome, KEGG and Biocarta pathways).

5 **d. Single and Pairs method**

In this method,

1. A threshold (a_0) is selected.
2. For each pair of genes the leave-one-out cross-validation of a quadratic discriminant analysis is computed and the accuracy recorded. If
10 this accuracy is less than a_0 , the pair of genes is saved.
3. For each gene, the a_0 and $1 - a_0$ quantile of the values in each class to discriminating the classes A and B, $Q_{a_0}(A)$, $Q_{1-a_0}(A)$, $Q_{a_0}(B)$, $Q_{1-a_0}(B)$, is computed. If either: $Q_{a_0}(A) > Q_{1-a_0}(B)$ or $Q_{1-a_0}(A) < Q_{a_0}(B)$, the gene is selected.
- 15 4. The obtained list is used to train M on the reduced feature space.
5. a_0 is chosen by cross-validation.

e. Forward Learning method

In this method,

1. Set IN to the empty list, choose N (typically 20, 100, 200).
- 20 For $n = 1, \dots, N$ do
 - a. For each gene, g, not in IN, compute a randomForest (ntree=500) on the subspace corresponding to $\{IN, g\}$, record the out-of-the bag true positive rates (TPr) and true negative rates (TNr) and compute the g-performance $\sqrt{TPr * TNr}$.
 - 25 b. Select the gene, g_{max} , for which the g-performance is maximum and add it to the list $IN := \{IN, g_{max}\}$.
 - c. Then train M on the subspace given by IN. N is chosen by cross-validation.

[0135] The union of the gene signatures extracted from the results of the
30 foregoing methods is considered as the larger gene signature. A weight is given to each of the genes in the union of the gene signature on the basis of the number of times it appears in a generated signature and the length of each generated gene

signature. Genes in Table 1 are those that appear in at least 3 of 5 lists generated. The genes in Table 1 that appear in at least 4 of the 5 lists generated are more predictive of lung cancer status than those appearing in 3 of the 5 lists generated.

[0136] While implementations of the invention have been particularly shown and
5 described with reference to specific examples, it should be understood by those skilled in the art that various changes in form and detail may be made therein without departing from the spirit and scope of the disclosure.

Table 1:

Gene Symbol	ENTREZ_GENE_ID	Gene Title	present / total lists
ZIC2	7546	Zic family member 2 (odd-paired homolog, Drosophila)	4/5
LOC100131262	100131262	hypothetical LOC100131262	4/5
CD83	9308	CD83 molecule	4/5
EML1	2009	echinoderm microtubule associated protein like 1	4/5
PAIP1	10605	poly(A) binding protein interacting protein 1	4/5
NIPBL	25836	Nipped-B homolog (Drosophila)	4/5
CREB3L1	90993	cAMP responsive element binding protein 3-like 1	4/5
SLC37A1	54020	solute carrier family 37 (glycerol-3-phosphate transporter), member 1	4/5
SFMBT2	57713	Scm-like with four mbt domains 2	4/5
ALDH3B1	221	aldehyde dehydrogenase 3 family, member B1	3/5
WNT5A	7474	wingless-type MMTV integration site family, member 5A	3/5
CALML3	810	calmodulin-like 3	3/5
SLC44A4	80736	solute carrier family 44, member 4	3/5
USP8	9101	ubiquitin specific peptidase 8	3/5
SLC41A2	84102	solute carrier family 41, member 2	3/5
CTSH	1512	cathepsin H	3/5
CSTA	1475	cystatin A (stefin A)	3/5
HNF1B	6928	HNF1 homeobox B	3/5
DSC3	1825	desmocollin 3	3/5
C18ORF1	753	chromosome 18 open reading frame 1	3/5
C8ORF40	114926	chromosome 8 open reading frame 40	3/5
SLC25A40	55972	solute carrier family 25, member 40	3/5
PTPLB	201562	protein tyrosine phosphatase-like (proline instead of catalytic arginine), member b	3/5
RORC	6097	RAR-related orphan receptor C	3/5
DENND2C	163259	DENN/MADD domain containing 2C	3/5
USP1	7398	ubiquitin specific peptidase 1	3/5
FMOD	2331	fibromodulin	3/5
FOSB	2354	FBJ murine osteosarcoma viral oncogene homolog B	3/5
PRAF2	11230	PRA1 domain family, member 2	3/5
IL1R2	7850	interleukin 1 receptor, type II	3/5
IER3IP1	51124	immediate early response 3 interacting protein 1	3/5
NSUN5	55695	NOP2/Sun domain family, member 5	3/5
DOK5	55816	docking protein 5	3/5

Gene Symbol	ENTREZ_GENE_ID	Gene Title	present / total lists
ECHDC2	55268	enoyl CoA hydratase domain containing 2	3/5
FKBP11	51303	FK506 binding protein 11, 19 kDa	3/5
THG1L	54974	tRNA-histidine guanylyltransferase 1-like (S. cerevisiae)	3/5
AVEN	57099	apoptosis, caspase activation inhibitor	3/5
KCNK5	8645	potassium channel, subfamily K, member 5	3/5
C9ORF167	54863	chromosome 9 open reading frame 167	3/5
TMEM117	84216	transmembrane protein 117	3/5
LIFR	3977	leukemia inhibitory factor receptor alpha	3/5
MLL5	55904	myeloid/lymphoid or mixed-lineage leukemia 5 (trithorax homolog, Drosophila)	3/5
MAGI3	260425	membrane associated guanylate kinase, WW and PDZ domain containing 3	3/5
SIKE1	80143	suppressor of IKBKE 1	3/5
LOC100507153	100507153	hypothetical LOC100507153	3/5
DLEU2	8847	deleted in lymphocytic leukemia 2 (non-protein coding)	3/5
LOC100132832	100132832	postmeiotic segregation increased 2-like 5-like	3/5
CCND2	894	cyclin D2	3/5
MYL9	10398	myosin, light chain 9, regulatory	3/5
MMP2	4313	Matrix metalloproteinase 2 (gelatinase A, 72kDa gelatinase, 72kDa type IV collagenase)	3/5
HTRA1	5654	HtrA serine peptidase 1	3/5
IGBP1	3476	immunoglobulin (CD79A) binding protein 1	3/5
ACTG2	72	actin, gamma 2, smooth muscle, enteric	3/5
TACC2	10579	transforming, acidic coiled-coil containing protein 2	3/5
MYO1E	4643	myosin IE	3/5
FBLN2	2199	fibulin 2	3/5
SRSF10	10772	serine/arginine-rich splicing factor 10	3/5
POT1	25913	protection of telomeres 1 homolog (S. pombe)	3/5
PSPH	5723	phosphoserine phosphatase	3/5
PDGFRL	5157	platelet-derived growth factor receptor-like	3/5
GOLGA8A	23015	golgin A8 family, member A	3/5
TST	7263	thiosulfate sulfurtransferase (rhodanese)	3/5
FAP	2191	fibroblast activation protein, alpha	3/5
DDX39B	10212	DEAD (Asp-Glu-Ala-Asp) box polypeptide 39	3/5
BTG3	10950	BTG family, member 3	3/5
DUSP7	1849	dual specificity phosphatase 7	3/5
RABGAP1L	9910	RAB GTPase activating protein 1-like	3/5
ZYX	7791	zyxin	3/5

Gene Symbol	ENTREZ_GENE_ID	Gene Title	present / total lists
SLC7A8	23428	solute carrier family 7 (amino acid transporter, L-type), member 8	3/5
IDE	3416	insulin-degrading enzyme	3/5
WFDC1	58189	WAP four-disulfide core domain 1	3/5
KDELC1	79070	KDEL (Lys-Asp-Glu-Leu) containing 1	3/5
LINC00339			3/5
KCNE4	23704	potassium voltage-gated channel, Isk-related family, member 4	3/5
HORMAD1	84072	HORMA domain containing 1	3/5
CDCA7L	55536	cell division cycle associated 7-like	3/5
CTHRC1	115908	collagen triple helix repeat containing 1	3/5
HIP1	3092	huntingtin interacting protein 1	3/5
ATP6V1C1	528	ATPase, H ⁺ transporting, lysosomal 42kDa, V1 subunit C1	3/5
ZNF521	25925	zinc finger protein 521	3/5
LOC100131564	100131564	hypothetical LOC100131564	3/5
ERP27	121506	endoplasmic reticulum protein 27	3/5
IGDCC4	57722	immunoglobulin superfamily, DCC subclass, member 4	3/5
SCAI	286205	suppressor of cancer cell invasion	3/5
FAM165B	54065	family with sequence similarity 165, member B	3/5
ALG10B	144245	asparagine-linked glycosylation 10, alpha-1,2-glucosyltransferase homolog B (yeast)	3/5
ZNF92	168374	zinc finger protein 92	3/5

5 We claim:

1. A method of classifying or grading a lung cancer tumor in an individual at risk for or having lung cancer comprising detecting the expression level of at least 2 genes listed in Table 1 in a test sample obtained from the individual; wherein a differential pattern of expression levels of said at least 2 genes in the test sample
10 classifies the lung cancer tumor as one of stage 1 lung adenocarcinoma, stage 2 adenocarcinoma, stage 1 squamous cell carcinoma or stage 2 squamous cell carcinoma.
2. The method according to claim 1, wherein the differential pattern of
15 expression levels is identified by a classifier based on a plurality of genes listed in Table 1, including said at least two genes, said classifier having been trained by *in silico* analysis or one or more feature selection and classification algorithms.
3. The method according to claim 1 or 2, wherein the differential pattern of
20 expression levels is identified by a classifier based on a plurality of genes listed in Table 1, including said at least two genes, said classifier having been trained by one or more algorithms selected from the group consisting of dual ensemble, generalized simulated annealing, T-filter, CORG, CORG combined with support vector machine, dual bagging, single and pairs, forward learning, Laplacian based
25 learning and learning method based on network perturbation amplitude.
4. The method according to any one of claims 1-3, wherein said classifier having been trained with at least the data in the Gene Expression Omnibus datasets GSE2109, GSE10245, GSE18842 and GSE37745.
30
5. The method according to any one of claims 1-4, wherein the method further comprises comparing the expression level of said at least 2 genes in the test sample and a control sample; or detecting the expression level of said at least 2 genes in

the control sample and comparing the expression level of said at least 2 genes in
35 the test sample and control sample, to identify the differential pattern.

6. The method according to any one of claims 1-5, wherein said at least 2
genes are selected from the group consisting of: ZIC2, LOC100131262, CD83,
EML1, PAIP1, NIPBL, CREB3L1, SLC37A1, and SFMBT2.

40

7. The method according to any one of claims 1-6, wherein the expression
level of said at least 2 genes in the test sample are detected by using a human
genome-wide array, a human lung tissue array or a custom array comprising
polynucleotides of a plurality of genes in Table 1 and said at least 2 genes.

45

8. The method according to any one of claims 1-6, wherein the expression
level of said at least 2 genes in the test sample are detected by measuring the level
of proteins encoded by the genes.

50 9. An array comprising polynucleotides hybridizing to at least 2 lung cancer
signature genes immobilized on a solid surface, wherein the lung cancer signature
genes are selected from the genes listed in Table 1 and said array is not a human
genome-wide array.

55 10. A device comprising antibodies immobilized on a solid surface that bind to
proteins encoded by at least 2 lung cancer signature genes, wherein the lung cancer
signature genes are selected from the group consisting of the genes listed in
Table 1.

60 11. A computer readable medium or computer program product comprising a
classifier based on at least two genes listed in Table 1, said classifier having been
trained by *in silico* analysis or one or more feature selection and classification
algorithms.

65 12. The computer readable medium or computer program product according to
claim 11, wherein said classifier is trained by one or more algorithms selected from
the group consisting of dual ensemble, generalized simulated annealing, T-filter,
CORG, CORG combined with support vector machine, dual bagging, single and
70 pairs, forward learning, Laplacian based learning and learning method based on
network perturbation amplitude.

13. The computer readable medium or computer program product according to
claim 11 or 12, wherein said classifier is trained with at least the data in the Gene
Expression Omnibus datasets GSE2109, GSE10245, GSE18842 and GSE37745.

75

14. The computer readable medium or computer program product according to
any one of claims 11-13, wherein said at least two genes are selected from the
group consisting of ZIC2, LOC100131262, CD83, EML1, PAIP1, NIPBL,
CREB3L1, SLC37A1, and SFMBT2.

80

15. A kit for classifying and grading a lung cancer tumor or for assessing the
prognosis of lung cancer in an individual, comprising one or more reagents that
detects expression levels of at least 2 genes listed in Table 1 in a test sample and
instructions for using said kit for classifying and grading a lung cancer tumor or for
85 determining the prognosis of lung cancer in said individual.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/EP2013/062993

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:
2. ☒ Claims Nos.: 9, 10, 15(all partially)
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
see FURTHER INFORMATION sheet PCT/ISA/210
3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of Item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- ☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- ☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2013/062993

A. CLASSIFICATION OF SUBJECT MATTER
INV. C12Q1/68
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, BIOSIS, EMBASE, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2009/067546 A2 (CELERA CORPRATION [US]; BIRSE CHARLES [US]; RUBEN STEVE [US]; LEWIS MA) 28 May 2009 (2009-05-28) claims 9-11; tables 7, 8 -----	1-10
X	WO 2009/146545 A1 (UNIV HEALTH NETWORK [CA]; TSAO MING-SOUND [CA]; AVIEL-RONEN SARIT [IL]) 10 December 2009 (2009-12-10) paragraph [0124]; claim 75 ----- -/--	9,10



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

5 September 2013

Date of mailing of the international search report

16/09/2013

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Santagati, Fabio

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2013/062993

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	L. GUO ET AL: "Constructing Molecular Classifiers for the Accurate Prognosis of Lung Adenocarcinoma", CLINICAL CANCER RESEARCH, vol. 12, no. 11, 1 June 2006 (2006-06-01), pages 3344-3354, XP055077626, ISSN: 1078-0432, DOI: 10.1158/1078-0432.CCR-05-2336 table 2	1-10
X	----- M. ROTUNNO ET AL: "A Gene Expression Signature from Peripheral Whole Blood for Stage I Lung Adenocarcinoma", CANCER PREVENTION RESEARCH, vol. 4, no. 10, 1 October 2011 (2011-10-01), pages 1599-1608, XP055077722, ISSN: 1940-6207, DOI: 10.1158/1940-6207.CAPR-10-0170 the whole document -----	1-10

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/EP2013/062993

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2009067546 A2	28-05-2009	CA 2705486 A1 EP 2227556 A2 JP 2011527414 A US 2009176228 A1 US 2012076725 A1 WO 2009067546 A2	28-05-2009 15-09-2010 27-10-2011 09-07-2009 29-03-2012 28-05-2009
WO 2009146545 A1	10-12-2009	CA 2726531 A1 US 2011224313 A1 US 2013065789 A1 WO 2009146545 A1	10-12-2009 15-09-2011 14-03-2013 10-12-2009

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

Continuation of Box II.2

Claims Nos.: 9, 10, 15(all partially)

Present claims 9, 10 and 15 relate to an extremely large number of possible products deriving from all the possible combinations of any two genes of Table 1. No support and disclosure in the sense of Article 6 and 5 PCT is to be found however for any combination, apart from the whole set of genes of Table 1 and the minimum set of the first nine most significant genes.

The non-compliance with the substantive provisions is to such an extent, that the search was performed taking into consideration the non-compliance in determining the extent of the search (PCT Guidelines 9.19 and 9.23).

The search of claims 9, 10 and 15 was restricted so as to include the whole set of genes of Table 1 or the top nine genes of the same table.

The applicant's attention is drawn to the fact that claims relating to inventions in respect of which no international search report has been established need not be the subject of an international preliminary examination (Rule 66.1(e) PCT). The applicant is advised that the EPO policy when acting as an International Preliminary Examining Authority is normally not to carry out a preliminary examination on matter which has not been searched. This is the case irrespective of whether or not the claims are amended following receipt of the search report or during any Chapter II procedure. If the application proceeds into the regional phase before the EPO, the applicant is reminded that a search may be carried out during examination before the EPO (see EPO Guidelines C-IV, 7.2), should the problems which led to the Article 17(2) declaration be overcome.