



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

(21) International Application Number:
PCT/EP2013/062996

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

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/662,779 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-2034 Peseux (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 COPD DIAGNOSIS

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



WO 2013/190092 A1

Gene Signatures for COPD Diagnosis

5 Field of the invention

[0001] The present invention relates to gene signatures that are useful for the diagnosis of Chronic Obstructive Pulmonary Disease (COPD). The present invention also relates to methods of diagnosing COPD. The invention further relates to arrays and computer readable media comprising gene signatures for
10 diagnosis of COPD.

Background of the Invention

[0002] Chronic Obstructive Pulmonary Disease (COPD) is the third leading cause of death in the U.S. and is projected to be the fourth leading cause of death worldwide by 2030. In Europe, over €20 billion is spent treating COPD, while
15 approximately \$18 billion is spent in the U.S.

[0003] COPD is a progressive disease that affects an individual's ability to breathe over time and includes two main conditions: emphysema and chronic bronchitis. In COPD, the airways in the lung chronically narrow, limiting the amount of airflow through the airways. Individuals affected by COPD experience
20 decreased elasticity in their airways and air sacs, the destruction of the walls between air sacs, the inflammation of the walls of the airways, and the clogging of airways by increased production of mucus.

[0004] COPD is caused by long-term exposure to lung irritants in the environment, such as air pollution, chemical fumes, dust and tobacco smoke.
25 Because long-term exposure to lung irritants is typically required, most individuals who suffer from COPD are more than 40 years old when symptoms first present. Such symptoms include an ongoing cough or a cough that produces an excess of mucus; shortness of breath, especially with physical activity; wheezing and chest tightness. COPD symptoms slowly worsen over time, and most affected

individuals do not notice symptoms at first because they are so mild or easy to correct by lifestyle adjustment.

[0005] COPD is typically diagnosed by signs and symptoms, including medical history, family history and test results. Diagnostic tests for COPD include lung function tests, such as spirometry, lung volume measurement, and lung diffusion capacity; chest X-rays, chest CT scans and arteriole blood gas tests. Accordingly, the diagnostic tests for COPD require the disease to have progressed to the point that lung function is moderately affected. Thus, there is a need for a diagnostic test that can identify COPD in patients at early stages. Recently, there have been attempts to identify genes associated with COPD or the treatment thereof. *See, e.g.,* US 2011/0160070; US 2010/0119474; and US 2009/0186951. A need still persists to understand the molecular mechanisms of COPD, which may allow for the design or optimization of therapies to treat the disease, instead of just the symptoms.

15 Summary of the invention

[0006] The present invention is directed to gene signatures for classifying, diagnosing or grading Chronic Obstructive Pulmonary Disease (COPD). A first aspect of the invention provides a method of diagnosing, classifying or grading COPD in an individual at risk for or having COPD. In some embodiments the method comprises classifying a test sample as COPD or non-COPD. In some 20 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 COPD or non-COPD. In some embodiments, a differential pattern of 25 expression levels of said at least 2 genes in the test sample diagnoses, classifies or grades the COPD.

[0007] 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 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 30

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
5 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 GSE10106, GSE10135, GSE11906, GSE11952, GSE13933, GSE19407, GSE19667, GSE20257, GSE5058, GSE7832, and GSE8545.

10 **[0008]** In some embodiments, the method comprises detecting the expression level of at least 2 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
15 1 is different in the test sample than in the control sample, then the individual suffers from COPD. In some embodiments, the expression level of the genes listed in Table 2 is higher in the test sample than in the control sample. Optionally, the expression level of the genes listed in Table 3 is lower in the test sample than in the control sample. In some embodiments, the method further comprises detecting the
20 expression level of at least 2 genes listed in Table 1 in the control sample.

[0009] In some embodiments, the at least 2 genes are selected from the group consisting of: PROS1, IRAK1, VAV3, FUT3, SFN, ZBTB44, CLDN8, BMPR1A, PAPD4, VCL, PPP2R5C, DGKA, and CYP51A1.

[0010] In some embodiments, the test sample or control sample is selected from
25 blood, serum, plasma, sputum, saliva, tissue, bronchia brushings, exhaled breath, and urine. Optionally, the test sample is obtained from a large airway of the individual, such as from a bronchial brush inserted into the large airway of the individual. In some embodiments, the control sample is obtained from a large airway of an individual not affected with COPD, such as from a bronchial brush
30 inserted into the large airway of the individual not affected with COPD. In some embodiments, the control sample is obtained from the individual at risk for or

having COPD prior to onset of COPD. In other embodiments, the control sample is obtained from an individual that does not suffer from COPD.

[0011] 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 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.

[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 the level of proteins encoded by the 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 both mRNA levels and the level of proteins encoded by the 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 biological sample are compared by *in silico* analysis (e.g., network-based analysis or machine-learning methods).

[0015] A second aspect of the invention provides an array for use in diagnosing or prognosing COPD. In some embodiments, the array comprises polynucleotides immobilized on a solid surface that can hybridize to at least 2 signature genes, wherein the COPD signature genes are selected from the group consisting of the genes listed in Table 1. Optionally, the array 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.

[0016] A third aspect of the invention provides a panel for use in diagnosing or prognosing COPD. In some embodiments, the panel comprises antibodies immobilized on a solid surface that bind to proteins encoded by at least 2 COPD

signature genes, wherein the COPD signature genes are selected from the group consisting of the genes listed in Table 1.

[0017] A fourth aspect of the invention provides a computer readable medium or a computer program for use in diagnosing or prognosing COPD. In some
5 embodiments, the computer readable medium or computer program comprises a COPD gene signature, wherein the gene signature comprises at least 2 genes selected from the genes listed in Table 1.

[0018] In some embodiments, the computer readable medium or computer
program product comprises a classifier based on at least two genes listed in Table
10 1, said classifier having been trained by in silico analysis or one or more feature 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
15 learning and learning method based on network perturbation amplitude. The classifier may be trained with at least the data in the Gene Expression Omnibus datasets GSE10106, GSE10135, GSE11906, GSE11952, GSE13933, GSE19407, GSE19667, GSE20257, GSE5058, GSE7832, and GSE8545. In some
embodiments, the at least two genes are selected from the group consisting of
20 PROS1, IRAK1, VAV3, FUT3, SFN, ZBTB44, CLDN8, BMPR1A, PAPD4, VCL, PPP2R5C, DGKA, and CYP51A1.

[0019] A fifth aspect of the invention provides a device for diagnosing or prognosing COPD. In some embodiments, the device comprises means for detecting the expression level of the genes listed in Table 1 in a test sample; means
25 for correlating the expression level with a diagnosis or prognosis of the COPD; and means for outputting the COPD diagnosis or prognosis. Optionally, the device further comprises means for detecting the expression level of the genes listed in Table 1 in a control sample.

[0020] A sixth aspect of the invention provides a kit for diagnosing or
30 prognosing COPD. 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 diagnosing the COPD. In other embodiments, the

kit is for assessing the prognosis of COPD 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 COPD in said individual. In some
5 embodiments, the set of reagents that detects 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.

[0021] An seventh aspect of the invention provides a method of diagnosing COPD in an individual or of assessing the prognosis of an individual with COPD.

10 In some embodiments, the method comprises 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; 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
15 predictive of the diagnosis of COPD in the individual. In some embodiments, the method comprises 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; calculating a numerical biomarker score for the individual based on the expression levels of the biomarkers measured in step
20 a); wherein the numerical biomarker score is predictive of the prognosis of COPD in the individual.

[0022] 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
25 a tumor.

[0023] 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
30 comprising polynucleotides of a plurality of genes in Table 1 and said at least 2 genes.

[0024] 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.

[0025] In some embodiments, the expression level of the genes listed in Table 1
5 in the biological sample is detected by measuring both mRNA levels and 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 machine-learning methods.

10 [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 diagnosing an individual as being at risk for or having Chronic
15 Obstructive Pulmonary Disease (COPD) 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,
20 wherein, if the expression level of said at least 2 genes is different in the test sample than in the control sample, then the individual suffers from COPD.
2. The method according to paragraph 1, wherein the expression level of the genes listed in Table 2 is higher in the test sample than in the control sample.
3. The method according to paragraph 1 or 2, wherein the expression level of
25 the genes listed in Table 3 is lower in the test sample than in the control sample.
4. The method according to any one of paragraphs 1-3, wherein the method further comprises detecting the expression level of said at least 2 genes in the control sample.
5. The method according to any one of paragraphs 1-4, wherein the test
30 sample is selected from blood, serum, plasma, sputum, saliva, tissue, bronchia brushings, exhaled breath, and urine.

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

18. The method according to any one of paragraphs 1-11, wherein the expression level of said at least 2 genes in the test sample and the expression level of said at least 2 genes in the control sample are detected by measuring the level of proteins encoded by the genes.
- 5 19. The method according to paragraph 18, wherein the protein level is measured using an antibody assay or by mass spectroscopy.
20. The method according to paragraph 19, wherein the antibody assay is selected from Western analysis, immunofluorescence, ELISA, and immunohistochemistry.
- 10 21. The method according to any one of paragraphs 1-20, 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.
22. The method according to paragraph 21, wherein the *in silico* analysis comprises using a classifier generated by one or more network-based methods or machine-learning based methods.
- 15 23. An array comprising polynucleotides hybridizing to at least 2 COPD signature genes immobilized on a solid surface, wherein the COPD signature genes are selected from the genes listed in Table 1.
- 20 24. A panel comprising antibodies immobilized on a solid surface that bind to proteins encoded by at least 2 COPD signature genes, wherein the COPD signature genes are selected from the genes listed in Table 1.
25. A computer readable medium comprising a gene signature, wherein the gene signature comprises at least 2 genes selected from the genes listed in Table 1.
- 25 26. A device for diagnosing COPD, the device comprising: 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 diagnosis of the COPD; and means for outputting the COPD diagnosis.
- 30 27. A device for prognosing COPD, the device comprising: 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 prognosis of the COPD; and means for outputting the COPD prognosis.

28. A kit for classifying and grading COPD, comprising a set of 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 COPD in said individual.

29. A kit for assessing the prognosis of COPD 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 the COPD in said individual.

30. A method of diagnosing, prognosing, classifying or grading COPD in a biological sample or an individual comprising measuring the expression levels of at least 2 genes listed in Table 1 in said biological sample or a test sample obtained from said 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 COPD or non-COPD.

31. The method according to paragraph 30, wherein a classifier or a previously established standard is used to determine whether a test sample is a COPD or non-COPD.

32. The method according to paragraph 31, wherein the classifier is obtained by training with a network-based method or a machine-learning based method using datasets obtained from subjects with COPD and datasets from subjects without COPD.

33. A method of diagnosing an individual as being at risk for or having Chronic Obstructive Pulmonary Disease (COPD) comprising detecting the expression level of at least 2 of the 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 diagnoses the individual as suffering from COPD.

34. The method according to paragraph 33, 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.

35. The method according to paragraph 33 or 34, 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,

5 Laplacian based learning and learning method based on network perturbation amplitude.

36. The method according to any one of paragraphs 33-35, wherein said classifier having been trained with at least the data in the Gene Expression Omnibus datasets GSE10106, GSE10135, GSE11906, GSE11952, GSE13933,
10 GSE19407, GSE19667, GSE20257, GSE5058, GSE7832, and GSE8545.

37. The method according to any one of paragraphs 33-36, 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
15 genes in the test sample and control sample, to identify the differential pattern.

38. The method according to any one of paragraphs 33-37, wherein said at least 2 genes are selected from the group consisting of: PROS1, IRAK1, VAV3, FUT3, SFN, ZBTB44, CLDN8, BMPR1A, PAPD4, VCL, PPP2R5C, DGKA, and CYP51A1.

20 39. The method according to any one of paragraphs 33-38, 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.

25 40. The method according to any one of paragraphs 33-38, 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.

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

42. A device comprising antibodies immobilized on a solid surface that bind to proteins encoded by at least 2 COPD signature genes, wherein the COPD signature genes are selected from the genes listed in Table 1.

43. 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.

44. The computer readable medium or computer program product according to paragraph 43, 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 pairs, forward learning, Laplacian based learning and learning method based on network perturbation amplitude.

45. The computer readable medium or computer program product according to paragraph 44, wherein said classifier is trained with at least the data in the Gene Expression Omnibus datasets GSE10106, GSE10135, GSE11906, GSE11952, GSE13933, GSE19407, GSE19667, GSE20257, GSE5058, GSE7832, and GSE8545.

46. The computer readable medium or computer program product according to any one of paragraphs 43-45, wherein said at least two genes are selected from the group consisting of PROS1, IRAK1, VAV3, FUT3, SFN, ZBTB44, CLDN8, BMPR1A, PAPD4, VCL, PPP2R5C, DGKA, and CYP51A1.

Brief Description of the Drawings

[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.

[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.

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

[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 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).

[0036] The term “Chronic Obstructive Pulmonary Disease” or “COPD” refers to a complex disease that results in progressive loss of lung function. COPD is typically characterized by the occurrence of chronic bronchitis or emphysema, both of which result in airway narrowing. Clinically, COPD is typically detected by low airflow in lung function tests. COPD is typically irreversible and gets progressively worse over time. Symptoms of COPD include chronic cough, chronic sputum production, dyspnea, rhonchi, wheezing, chest tightness, tiredness and decreased airflow in lung function tests. Individuals suffering from very severe COPD can develop respiratory failure and present with cyanosis, headaches, drowsiness, and/or asterixis. COPD is a progressive disease and prognosis of the

disease can be predicted by severe airflow obstruction, poor exercise capacity, shortness of breath, being significantly underweight or overweight, respiratory failure, cor pulmonale, and frequent acute exacerbations. COPD prognosis can be evaluated using the BODE index, which is a scoring system that measures FEV1, body-mass index, 6-minute walk distance, and a modified MRC (Medical Research Council) dyspnea scale to estimate outcomes in COPD.

[0037] The term "classifying COPD" refers to a method for determining the type of COPD from which a subject suffers. COPD can be classified as primarily bronchial or primarily emphysematous. Such classifications are made by simply analyzing clinical, functional, and radiological findings or by detecting biomarkers. In bronchial COPD, lung damage and inflammation occurs in the large airways resulting in chronic bronchitis, which is characterized by hyperplasia and hypertrophy of goblet cells and mucous glands in the airway. In emphysematous COPD, lung damage and inflammation occurs in the alveoli and is characterized by enlargement of the air spaces distal to the terminal bronchioles, with destruction of their walls and a reduction in lung elasticity. COPD may be classified based on the presence, absence, alteration or levels of biomarkers. COPD may be classified based on the COPD gene signature. Classifying COPD may also refer to distinguishing between bronchial or emphysematous COPD.

[0038] 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 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.

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 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 or more processors and engines that communicate through a network interface.

Such an implementation may be appropriate for distributed computing over multiple communication systems.

[0039] 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 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 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 FLASH-EEPROM, barcodes, semiconductors, microchips and any other memory chip or cartridge.

[0040] 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 sample from an individual suffering from COPD, including but not limited to, GOLD stage 1, GOLD stage 2, GOLD stage 3, or GOLD stage 4 COPD. 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

5 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

10 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 individual, or some other sample. For example, a control sample may be blood, blood cells, serum, plasma, sputum, saliva, tissue, bronchial washing,

15 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.

20 [0041] 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

25 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

30 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.

- [0042] The term “forced expiratory volume in one second” or “FEV1” refers to the volume of air that can forcibly be blown out in one second, after full inspiration. Average values for FEV1 in healthy individuals depend on sex and age and have been well-characterized in the art. FEV1 and the FEV1 to FVC ratio (FEV1/FVC) are used clinically to grade COPD. In healthy adults FEV1/FVC should be approximately 75–80%. In obstructive diseases, such as COPD, FEV1 is diminished because of increased airway resistance to expiratory flow. While the FVC may be decreased as well, due to the premature closure of airway in expiration, FEV1 is typically more affected because of the increased airway resistance, so the FEV1/FVC ratio reflects the degree of airway closure compared to lung volume.
- [0043] The term “forced vital capacity” or “FVC” refers to the volume of air that can forcibly be blown out after full inspiration, measured in liters.
- [0044] 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, a treated state or a recovery state. A gene signature may be characterized by which genes are expressed 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 and/or grading of COPD in an individual. The gene signature may be unique to the class and grade of COPD.
- [0045] The term “grading COPD” refers to a method for determining the grade of COPD from which a subject suffers. There are several different grades of COPD, which reflect the severity of the disease. For example, the Global Initiative for Chronic Obstructive Lung Disease (GOLD) characterizes COPD patients into GOLD Stage 1 to 4 depending on the severity of disease. GOLD stage 0 refers to a high risk population who did not present the symptoms used to describe stage 1; stage 1 refers to mild COPD and is characterized by a FEV1/FVC ratio less than 70% and an FEV1 greater than 80%; stage 2 refers to moderate COPD and is

characterized by a FEV1/FVC ratio less than 70% and an FEV1 between 50% and 80%; stage 3 refers to severe COPD and is characterized by a FEV1/FVC ratio less than 70% and an FEV1 between 30% and 50% and stage 4 refers to very severe COPD and is characterized by a FEV1/FVC ratio less than 70% and an FEV1 less than 30% or the presence of chronic renal failure or right heart failure.

[0046] 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 expression levels to obtain a classification of the test sample, e.g., as COPD or non-COPD. 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 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 application, incorporated herein by reference in its entirety, 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. 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.

[0047] 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.

[0048] The term “individual at risk for COPD” refers to an individual who is predisposed to COPD. Predisposition to COPD may be due to one or more genetic or environmental factors. For example, an individual related to a COPD patient is more likely to get COPD than an individual who is not related to a COPD patient.

5 Further, exposure to environmental factors such as radon gas, asbestos, tobacco smoke, and air pollution can increase the risk for COPD and predispose an individual to COPD.

[0049] The term “individual having COPD” or “individual suffering from COPD” refers to an individual experiencing progressive loss of lung function, typically characterized by airway narrowing. COPD can be bronchial or emphysematous and may be detected by analyzing clinical, functional, and radiological findings or detecting relevant biomarkers.

[0050] 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 via the time that it takes an ionized particle to reach a detector at a known distance.

[0051] 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.

[0052] The term “network-based methods” refers to methods for identifying 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.

[0053] 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.

[0054] 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.

[0055] 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.

[0056] 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.

[0057] The term “test sample” refers to a sample obtained from an individual at risk for, having or suffering from COPD. 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 COPD in the individual. A 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-fixed paraffin-embedded (FFPE)).

[0058] 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 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 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

[0059] One aspect of the invention provides gene signatures useful for
10 diagnosing, prognosing, classifying or grading COPD. 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 least 65, at least 70, at
15 least 75, at least 80 or at least 84 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, at least 5, or at least 10 genes are selected from the group consisting of: PROS1, IRAK1, VAV3, FUT3, SFN, ZBTB44, CLDN8, BMPRI1A, PAPD4, VCL, PPP2R5C, DGKA, and
20 CYP51A1, which are the genes that appear in 4 of the 5 lists generated in Example 1.

[0060] 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 or at least 35 of the genes selected from the genes listed in Table 2 have increased expression compared to a control
25 sample. In some embodiments, each of the genes listed in Table 2 has 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, or at least 45 of the genes selected from the genes listed in Table 3 have decreased expression compared to a control sample. In some
30 embodiments, each of the genes listed in Table 3 has decreased expression compared to a control sample.

[0061] 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.

[0062] The present invention encompasses the following gene signatures:

- i. A and B;
- 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;
- 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;
- 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;
- xvii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, and R;

- xxviii. A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, and S;
- xxix. 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;
- 5 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;
- 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;
- 10 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;
- 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;
- 15 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;
- 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;
- 20 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;
- 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;
- 25 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;
- 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;
- 30 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;

- 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;
- 5 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;
- 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;
- 20 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;
- 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;
- 30 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;

- 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,
15 AQ, AR, AS, AT, AU, AV, AW, AX, AY, AZ, BA and BB;
- 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,
20 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;
- 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,
30 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;

- 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;

- 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, BW, BX, BY, BZ, CA, CB, CC and CD;
- 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, BW, BX, BY, BZ, CA, CB, CC, CD and CE;
- 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, BW, BX, BY, BZ, CA, CB, CC, CD, CE and CF; and
- 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 and CG 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 and CG are different.
- [0063] In some embodiments, 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 and CG are independently selected from the genes listed in Table 2 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 and CG are different.

10 [0064] In some embodiments, 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 and CG are independently selected from the genes listed in Table 3 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 and CG are different.

Methods of Using Biomarkers and Gene Signatures

[0065] The biomarkers and gene signatures of the invention may be used in methods of diagnosing, prognosing, classifying or grading COPD in biological sample or an individual. One aspect of the invention provides a method of diagnosing, classifying or grading COPD in an individual at risk for or having COPD. In some embodiments the method comprises classifying a test sample as COPD or non-COPD. 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 COPD or non-COPD. 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 or at least 84 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
5 diagnoses, classifies or grades the COPD.

[0066] 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 a COPD sample from those datasets obtained from a non-COPD or healthy sample. In some methods of the invention, control data is
10 not collected or used. Instead, a classifier or a previously established standard may be used to determine whether a test sample is a COPD sample. For example, a classifier that is obtained by training with network-based or machine-learning based methods using datasets obtained from subjects with COPD and datasets from subjects without COPD can be used. Alternatively, one or more numerical scores
15 (e.g., average fold change or rank absolute value as shown in Table 1) 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 COPD sample or a normal sample.

20 [0067] In one embodiment, the invention provides a method of diagnosing COPD 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
25 method, or a combination of the foregoing methods. The classifier can be obtained from the network-based methods, the machine-learning based methods, or a combination of the foregoing methods by training with datasets obtained from subjects with COPD, subjects with COPD of a certain determined GOLD stage, or healthy subjects. In another embodiment, a classifier may be obtained, given
30 appropriate class(es) of training datasets, to identify a specific prognosis of the COPD, to indicate transition from a more severe or less severe GOLD stage or to

indicate that a particular treatment regimen should be used to treat the individual who provided the biological sample.

[0068] In one embodiment, the methods of the invention comprise obtaining a test sample (such as bronchial brushings) 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 COPD treatment regimen based on the comparison. In a further embodiment, the invention provides a method for monitoring the progress of a COPD 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.

[0069] 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.

[0070] In some embodiments, the method comprises detecting the expression level of at least 2 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 COPD. In some embodiments, the expression level of the genes listed in Table 2 is higher in the test sample than in the control sample. Optionally, the expression level of the genes listed in Table 3 is lower in the test sample than in the control sample. 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 or at least 84 genes listed in Table 1 are detected.

[0071] In some embodiments, the test sample is selected from blood, serum, plasma, sputum, saliva, tissue, bronchia brushings, exhaled breath, and urine. Optionally, the test sample is obtained from a large airway of the individual, such as from a bronchial brush inserted into the large airway of the individual.

[0072] In some embodiments, the control sample is selected from blood, serum, plasma, sputum, saliva, tissue, bronchia brushings, exhaled breath, and urine. Optionally, the control sample is obtained from a large airway of an individual not affected with COPD, such as from a bronchial brush inserted into the large airway of the individual not affected with COPD. In some embodiments, the control sample is obtained from the individual at risk for or having COPD prior to onset of COPD. In other embodiments, the control sample is obtained from an individual that does not suffer from COPD.

[0073] 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.

[0074] 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.

[0075] 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.

[0076] 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.

Methods of Biomarker Detection, Arrays and Panels

[0077] 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.

[0078] 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.

[0079] 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 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 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).

[0080] For example, these techniques find application in microarray-based assays 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 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.

[0081] 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.

[0082] 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, Table 2, or Table 3.

[0083] 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.

[0084] 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

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, 751-1000, or 1001-1500 bases in length.

[0085] 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 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 direction of transcription.

[0086] 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.

[0087] 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 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

- 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.
- 5 **[0088]** In one aspect of the invention, biomarkers may be detected in the test 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.
- 10 The cDNA may be used in a qPCR assay to produce fluorescence as the DNA 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.
- 15 Further details are provided, for example, in "Gene Expression Profiling: Methods and Protocols," Richard A. Shimkets, editor, Humana Press, 2004 and US patent application 2010/0070191.
- 20 **[0089]** The invention encompasses an array comprising polynucleotides that hybridize to genes listed in Table 1, Table 2, or Table 3. 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 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, Table 2, or Table 3. 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 embodiments, the array is not a human genome-wide array. Examples of human genome-wide array include, but are not limited to,
- 25 Exon 1.0 ST, Gene 1.0 ST, U 95, U133, U133A 2.0, and U133 Plus 2.
- 30 **[0090]** In another aspect of the invention, detection of the biomarkers described herein may be accomplished by an immunoassay procedure. The immunoassay

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 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 contacted with the test sample or the control sample.

[0091] 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 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* (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.

[0092] 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 Tables 1, 2, and 3.

[0093] 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.

[0094] 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 ¹²⁵I, ¹³¹I, ³⁵S or ³H.

[0095] Immunoassays for biomarker polypeptides will typically comprise incubating a sample, such as a biological fluid, a tissue extract, freshly harvested 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 (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 substrate in such a manner as to produce a chemical moiety which can be detected, for example, by spectrophotometric, or fluorimetric means.

[0096] 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 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.

[0097] 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.

[0098] Antibodies and other reagents may also be used to detect post-translational modifications (e.g., methylation, acetylation, farnesylation, biotinylation, stearylation, formylation, myristoylation, palmitoylation, geranylgeranylation, pegylation, phosphorylation, sulphation, glycosylation, sugar modification, lipidation, lipid modification, ubiquitination, sumolation, disulphide bonding, cysteinylation, oxidation, glutathionylation, carboxylation, glucuronidation, and deamidation) of biomarker proteins or biomarker polypeptides.

[0099] The invention encompasses a panel comprising antibodies that bind to proteins encoded by genes listed in Table 1, Table 2 or Table 3. 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, 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, Table 2 and Table 3. 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).

5 **[0100]** 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
10 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
15 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
20 information obtained from the detected ions into mass-to-charge ratios.

[0101] 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.

25 **[0102]** 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
30 fluorochrome-labeled antibodies that bind to one of the biomarker polypeptides of the invention.

[0103] 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
5 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
10 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 COPD.

In Silico Analysis and Computer Readable Media

15 [0104] Biomarkers and gene signatures of the invention may be predicted based on gene expression patterns in COPD. In some embodiments, biomarker and gene signature prediction comprises gene expression patterns in control (e.g., non-COPD) biological samples. A heterogeneous ensemble learning approach may be used to classify samples based on their gene expression profiles. Such an approach
20 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
25 approaches, disease module-based approaches, linkage methods, modularity or pathway-based methods and diffusion-based methods.

[0105] 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
30 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

5 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

10 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,

15 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, stearylolation, formylation, myristoylation, palmitoylation, geranylgeranylation, pegylation, phosphorylation, sulphation, glycosylation, sugar

20 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.

25 **[0106]** The test data sets may be processed and have their quality controlled together if they are obtained from the same technology platform (e.g., 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

30 R (R Development Core Team, *R: A Language and Environment for Statistical Computing*. 2007). The quality may be controlled by:

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;

5 2. Excluding arrays from the training datasets that fall below a set of 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
10 algorithm (Wu et al., *Journal of the American Statistical Association*, 99:909 (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
15 training samples.

[0107] 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
20 instant application and having the attorney docket no. 106500-0032-001), dual 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
25 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
30 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 is incorporated herein by reference in its entirety.

[0108] Generalized simulated annealing may be modified for binary functions. In one embodiment, a dual binary generalized simulated annealing based method
 5 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
 10 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
 15 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)).

[0109] The single and pairs method may include the following steps:

1. Select a threshold a_0 .
- 20 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, $Q_{a_0}(A)$, $Q_{1-a_0}(A)$, $Q_{a_0}(B)$, $Q_{1-a_0}(B)$.
 25 $Q_{a_0}(B)$. 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.

[0110] The forward learning method may include the following steps:

30 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 (TP_r) and true negative rates (TN_r) and compute the g-performance $\sqrt{TP_r * TN_r}$.
 - 5 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.
- [0111] The Laplacian based learning method may include the following steps:
- 10 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 $k=2,3,4,5$).
 - 15 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.
 - 20 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.
 - 25 7. Predict the test cases.
- [0112] 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.
- 30 [0113] 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

- (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, *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 reference in its entirety and having the attorney docket no. 106500-0022-001.
- [0114] 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.
- [0115] 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.
- [0116] 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.
- [0117] The genes obtained by these methods may be mapped to gene symbols using any suitable platform, for example, the *Confero* platform (Hermida et al., *Confero: an Integrated Contrast and Gene Set Platform for Computational Analysis and Biological Interpretation of Omics Data*. submitted, 2012).
- [0118] 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 embodied method.

[0119] The invention encompasses a method for classifying a test sample as a COPD sample or a non-COPD sample, the method comprising: measuring the
5 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 either a COPD sample or a non-COPD sample. In some embodiments, the expression levels of at least 2, at least 3, at
10 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 measured. In some embodiments, the classifier has been trained by *in silico* analysis or one or more feature selection and classification algorithms.

15 [0120] 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
20 information that identifies the sample. The computer readable medium may also include a computer program product.

[0121] 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
25 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.

[0122] 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
30 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 GSE10106, GSE10135, GSE11906, GSE11952, GSE13933, GSE19407, GSE19667, GSE20257, GSE5058, GSE7832, and GSE8545.

5 Devices and Kits

[0123] 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 COPD. The devices can comprise means for detecting the expression level of at least 2 genes listed in Table 1 or the level of at least 2
10 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
15 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 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
20 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 or at least 84 genes listed in Table 1. In some embodiments, the devices can comprise means for detecting the expression levels of 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,
25 at least 55, at least 60, at least 65, at least 70, at least 75, at least 80 or at least 84 genes listed in Table 1.

[0124] 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
30 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.

[0125] 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.

[0126] The devices may comprise a means for correlating the expression levels of the genes being analyzed with COPD 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 COPD or is suffering from stage 1, stage 2, stage 3 or stage 4 COPD. The models and/or algorithms can be provided in computer readable format.

[0127] The devices may also comprise output means for outputting the COPD 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 system for performing one or more of the steps provided.

[0128] Another aspect of the invention encompasses kits for practicing the methods of the invention. Such kits may be used for diagnosing, classifying and grading COPD or for assessing the prognosis of COPD in an individual. The kits can be used for clinical diagnosis and/or laboratory research. In one embodiment, a kit comprises in one or more containers one or more reagents that detect expression levels of genes that serve as biomarkers of COPD 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.

[0129] 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 COPD 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.

[0130] For protein-based methods, such as immunoassays, a detection reagent in the kit may comprise a biomarker antibody, which may be labeled or labelable. The antibodies may bind to proteins encoded by the COPD 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, 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 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 fluorochrome-labeled secondary antibodies. The kit may also include unlabeled or labeled antibodies to various cell surface antigens which can be used for identification or sorting of subpopulations of cells.

[0131] The detection reagents may be labeled or labelable by one or more detectable labels. Examples of detectable labels include, without limitation, 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).

[0132] 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.

[0133] 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.

Example 1 Generation of COPD Gene Signatures

[0134] 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 distinguishing COPD from healthy samples. 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.

[0135] 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.

The following public datasets are downloaded from the Gene Expression Omnibus

5 (GEO) (<http://www.ncbi.nlm.nih.gov/geo/>) repository:

- a. GSE10106 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE10106)
- b. GSE10135 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE10135)
- c. GSE11906 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE11906)
- d. GSE11952 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE11952)
- 10 e. GSE13933 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13933)
- f. GSE19407 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19407)
- g. GSE19667 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19667)
- h. GSE20257 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20257)
- i. GSE5058 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5058)
- 15 j. GSE7832 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7832)
- k. GSE8545 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE8545).

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
20 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;
- 25 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 gcma algorithm (Wu, 2004). Training set sample classifications are obtained from the
30 series matrix file of the GEO database for each dataset.

Arrays GSM101108.CEL, GSM252820.CEL, GSM252822.CEL,
GSM252824.CEL, GSM252826.CEL, GSM298246.CEL, GSM252819.CEL,

GSM252821.CEL, GSM252842.CEL, GSM252843.CEL, GSM252844.CEL,
 GSM252845.CEL, GSM252846.CEL, GSM252847.CEL, GSM252848.CEL,
 GSM252849.CEL, GSM252850.CEL, GSM252851.CEL, GSM252852.CEL,
 GSM252853.CEL, GSM252854.CEL, GSM114096.CEL, GSM114098.CEL,
 5 GSM114099.CEL, GSM114100.CEL, GSM114101.CEL, GSM114102.CEL,
 GSM114103.CEL, GSM114104.CEL, GSM114105.CEL, GSM114091.CEL,
 GSM300871.CEL, GSM300872.CEL, GSM300873.CEL, GSM300874.CEL,
 GSM300875.CEL, GSM300876.CEL, GSM300877.CEL, GSM300878.CEL,
 GSM300879.CEL, GSM300880.CEL, GSM491043.CEL, GSM491044.CEL,
 10 GSM300862.CEL, GSM300863.CEL, GSM300864.CEL, GSM300865.CEL,
 GSM300866.CEL, GSM300867.CEL, GSM300868.CEL, GSM300869.CEL,
 GSM300870.CEL, GSM252856.CEL, GSM252857.CEL, GSM252861.CEL,
 GSM252872.CEL, GSM252864.CEL, GSM101111.CEL, GSM252865.CEL,
 GSM252874.CEL, GSM252866.CEL, GSM190152.CEL, GSM252862.CEL,
 15 GSM252863.CEL, GSM252860.CEL, GSM252858.CEL, GSM252855.CEL,
 GSM252870.CEL, GSM252832.CEL, GSM252875.CEL, GSM252829.CEL,
 GSM252873.CEL, GSM101109.CEL, GSM434064.CEL were not used for further
 analysis. The output of this part consists of a gene expression matrix X on 273
 samples (233 training samples and 40 test samples) and 54675 probesets, and the
 20 class 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 analyzed to
 obtain a number of classifications for each test sample. Each method has defined
 25 input and 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.

30 **[0136]** 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.

5 [0137] Cross validation of the results is performed using any of the following supervised methods

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

[0138] The following methods are used to generate predictions:

a. Dual Ensemble method.

15 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
20 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

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)
25 automatically if the list size is over N. Any M is trained on the resulting subspace. N is chosen by cross-validation.

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
30 gene sets collection from the Broad Institute (Cambridge, MA) (Reactome, KEGG and Biocarta pathways).

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 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.
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).
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 M on the subspace given by IN. N is chosen by cross-validation.

[0139] The union of the gene signatures extracted from the results of the 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. A further reduction of the number of genes in the signature can be obtained by selecting the genes which has a fold-change value greater than a threshold (see Table 1, the column "fc>2" wherein a value of 1 is

ascribed to those genes that has a fold change value greater than 2). Further, 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.

- 5 **[0140]** While implementations of the invention have been particularly shown and 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: Genes which are specifically regulated in airway brushings of COPD patients compared with control airway brushings

probeset_ids	Gene Symbol	Entrez Gene Id	Gene Title	present / total lists	fold change in training (COPD vs control)	Rank by abs tvalue	direction of change in COPD vs. control	fc > 2
207808_s_at	PROS1	5627	protein S (alpha)	4/5	-2.3	64	downregulated	1
201587_s_at	IRAK1	3654	interleukin-1 receptor-associated kinase 1	4/5	1.2	2166	upregulated	0
218807_at	VAV3	10451	vav 3 guanine nucleotide exchange factor	4/5	-1.3	2636	downregulated	0
214088_s_at	FUT3	2525	fucosyltransferase 3 (galactoside 3(4)-L-fucosyltransferase, Lewis blood group)	4/5	1.4	11214	upregulated	0
33323_r_at	SFN	2810	stratifin	4/5	1.4	11226	upregulated	0
226148_at	ZBTB44	29068	zinc finger and BTB domain containing 44	4/5	-1.1	12186	downregulated	0
214598_at	CLDN8	9073	claudin 8	4/5	-1.4	14469	downregulated	0
213578_at	BMPRI1A	657	bone morphogenetic protein receptor, type 1A	4/5	-1.1	16909	downregulated	0
225761_at	PAPD4	167153	PAP associated domain containing 4	4/5	-1.1	17004	downregulated	0
200931_s_at	VCL	7414	vinculin	4/5	1.2	20026	upregulated	0
201877_s_at	PPP2R5C	5527	protein phosphatase 2, regulatory subunit B', gamma	4/5	-1.1	25186	downregulated	0
203385_at	DGKA	1606	diacylglycerol kinase, alpha 80kDa	4/5	1.1	29676	upregulated	0
202314_at	CYP51A1	1595	cytochrome P450, family 51, subfamily A, polypeptide 1	4/5	1.0	54179	upregulated	0

probeset_ids	Gene Symbol	Entrez Gene Id	Gene Title	present / total lists	fold change in training (COPD vs control)	Rank by abs tvalue	direction of change in COPD vs. control	fc > 2
231126_at	C2orf70	339778	chromosome 2 open reading frame 70	3/5	2.1	3	upregulated	1
206595_at	CST6	1474	cystatin E/M	3/5	2.9	50	upregulated	1
218130_at	C17orf62	79415	chromosome 17 open reading frame 62	3/5	1.4	52	upregulated	0
1556012_at	KLHDC7A	127707	kelch domain containing 7A	3/5	-1.9	87	downregulated	0
1552396_at	WFDC6	140870	WAP four-disulfide core domain 6	3/5	-2.5	101	downregulated	1
201761_at	MTHFD2	10797	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 2, methenyltetrahydrofolate cyclohydrolase	3/5	2.0	190	upregulated	1
221045_s_at	PER3	8863	period homolog 3 (Drosophila)	3/5	-1.7	242	downregulated	0
204369_at	PIK3CA	5290	phosphoinositide-3-kinase, catalytic, alpha polypeptide	3/5	-1.6	446	downregulated	0
210845_s_at	PLAUR	5329	plasminogen activator, urokinase receptor	3/5	2.6	467	upregulated	1
211596_s_at	LRIG1	26018	leucine-rich repeats and immunoglobulin-like domains 1	3/5	-1.3	484	downregulated	0
203221_at	TLE1	7088	transducin-like enhancer of split 1 (E(spl) homolog, Drosophila)	3/5	1.5	527	upregulated	0
201964_at	SETX	23064	senataxin	3/5	-1.4	533	downregulated	0
210397_at	DEFB1	1672	defensin, beta 1	3/5	2.6	591	upregulated	1
222791_at	RSBN1	54665	round spermatid basic protein 1	3/5	-1.5	605	downregulated	0

probeset_ids	Gene Symbol	Entrez Gene Id	Gene Title	present / total lists	fold change in training (COPD vs control)	Rank by absolute value	direction of change in COPD vs. control	fc > 2
202816_s_at	SS18	6760	synovial sarcoma translocation, chromosome 18	3/5	-1.4	733	downregulated	0
204106_at	TESK1	7016	testis-specific kinase 1	3/5	1.3	846	upregulated	0
207177_at	PTGFR	5737	prostaglandin F receptor (FP)	3/5	-1.7	891	downregulated	0
201666_at	TIMP1	7076	TIMP metalloproteinase inhibitor 1	3/5	1.5	1082	upregulated	0
209189_at	FOS	2353	FBJ murine osteosarcoma viral oncogene homolog	3/5	-4.1	1142	downregulated	1
204062_s_at	ULK2	9706	unc-51-like kinase 2 (C. elegans)	3/5	-1.4	1285	downregulated	0
209906_at	C3AR1	719	complement component 3a receptor 1	3/5	1.9	1568	upregulated	0
227194_at	FAM3B	54097	family with sequence similarity 3, member B	3/5	1.8	1874	upregulated	0
201884_at	CEACAM5	1048	carcinoembryonic antigen-related cell adhesion molecule 5	3/5	3.1	1971	upregulated	1
223805_at	OSBPL6	114880	oxysterol binding protein-like 6	3/5	-1.5	1976	downregulated	0
209369_at	ANXA3	306	annexin A3	3/5	1.5	2088	upregulated	0
222870_s_at	B3GNT2	10678	UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 2	3/5	-1.2	2429	downregulated	0
201041_s_at	DUSP1	1843	dual specificity phosphatase 1	3/5	-1.7	2469	downregulated	0
203510_at	MET	4233	met proto-oncogene (hepatocyte growth factor receptor)	3/5	-1.3	2537	downregulated	0
205513_at	TCN1	6947	transcobalamin I (vitamin B12 binding protein, R binder family)	3/5	3.4	2605	upregulated	1

probeset_ids	Gene Symbol	Entrez Gene Id	Gene Title	present / total lists	fold change in training (COPD vs control)	Rank by absolute value	direction of change in COPD vs. control	fc > 2
209022_at	STAG2	10735	stromal antigen 2	3/5	-1.2	4361	downregulated	0
201941_at	CPD	1362	carboxypeptidase D	3/5	-1.2	4845	downregulated	0
210058_at	MAPK13	5603	mitogen-activated protein kinase 13	3/5	1.3	5160	upregulated	0
201141_at	GNPMB	10457	glycoprotein (transmembrane) nmb	3/5	2.1	5266	upregulated	1
202539_s_at	HMGCR	3156	3-hydroxy-3-methylglutaryl-CoA reductase	3/5	-1.2	6340	downregulated	0
214070_s_at	ATP10B	23120	ATPase, class V, type 10B	3/5	1.5	8213	upregulated	0
226312_at	RICTOR	253260	RPTOR independent companion of MTOR, complex 2	3/5	-1.2	8387	downregulated	0
205019_s_at	VIPR1	7433	vasoactive intestinal peptide receptor 1	3/5	1.3	9695	upregulated	0
203182_s_at	SRPK2	6733	SRSF protein kinase 2	3/5	-1.1	10114	downregulated	0
210145_at	PLA2G4A	5321	phospholipase A2, group IVA (cytosolic, calcium-dependent)	3/5	1.4	10939	upregulated	0
205349_at	GNAI5	2769	guanine nucleotide binding protein (G protein), alpha 15 (Gq class)	3/5	1.2	11114	upregulated	0
213070_at	PIK3C2A	5286	phosphoinositide-3-kinase, class 2, alpha polypeptide	3/5	-1.2	11271	downregulated	0
209218_at	SQLE	6713	squalene epoxidase	3/5	-1.2	12374	downregulated	0
1558486_at	ZNF493	284443	zinc finger protein 493	3/5	-1.3	12655	downregulated	0

probeset_ids	Gene Symbol	Entrez Gene Id	Gene Title	present / total lists	fold change in training (COPD vs control)	Rank by absolute value	direction of change in COPD vs. control	$f_c > 2$
203655_at	XRCC1	7515	X-ray repair complementing defective repair in Chinese hamster cells 1	3/5	1.1	12911	upregulated	0
200985_s_at	CD59	966	CD59 molecule, complement regulatory protein	3/5	-1.1	13226	downregulated	0
208881_x_at	ID11	3422	isopentenyl-diphosphate delta isomerase 1	3/5	-1.2	13569	downregulated	0
205080_at	RARB	5915	retinoic acid receptor, beta	3/5	-1.2	13658	downregulated	0
212533_at	WEE1	7465	WEE1 homolog (S. pombe)	3/5	-1.1	14607	downregulated	0
219485_s_at	PSMD10	5716	proteasome (prosome, macropain) 26S subunit, non-ATPase, 10	3/5	-1.1	14730	downregulated	0
208820_at	PTK2	5747	PTK2 protein tyrosine kinase 2	3/5	-1.1	15909	downregulated	0
201384_s_at	NBR1	4077	Neighbor of BRCA1 gene 1	3/5	-1.1	17242	downregulated	0
201860_s_at	PLAT	5327	plasminogen activator, tissue	3/5	1.4	18596	upregulated	0
227038_at	SGMS2	166929	sphingomyelin synthase 2	3/5	-1.1	20534	downregulated	0
225240_s_at	MSI2	124540	musashi homolog 2 (Drosophila)	3/5	-1.1	20773	downregulated	0
207365_x_at	USP34	9736	ubiquitin specific peptidase 34	3/5	1.1	21072	upregulated	0
204971_at	CSTA	1475	cystatin A (stefin A)	3/5	1.3	21314	upregulated	0
203652_at	MAP3K11	4296	mitogen-activated protein kinase kinase kinase 11	3/5	1.2	22650	upregulated	0
202277_at	SPTLC1	10558	serine palmitoyltransferase, long chain base subunit 1	3/5	-1.1	22925	downregulated	0
229738_at	DNAH10	196385	dynein, axonemal, heavy chain 10	3/5	-1.1	24994	downregulated	0

- 66 -

probeset_ids	Gene Symbol	Entrez Gene Id	Gene Title	present / total lists	fold change in training (COPD vs control)	Rank by absolute value	direction of change in COPD vs. control	fc > 2
203060_s_at	PAPSS2	9060	3'-phosphoadenosine 5'-phosphosulfate synthase 2	3/5	1.2	25286	upregulated	0
202603_at	ADAM10	102	ADAM metalloproteinase domain 10	3/5	-1.1	27094	downregulated	0
201528_at	RPA1	6117	replication protein A1, 70kDa	3/5	-1.1	28648	downregulated	0
214484_s_at	SIGMAR1	10280	sigma non-opioid intracellular receptor 1	3/5	-1.1	31396	downregulated	0
227520_at	TXLNG	55787	taxilin gamma	3/5	-1.1	32441	downregulated	0
224370_s_at	CAPS2	84698	calyphosine 2	3/5	1.1	34524	upregulated	0
207556_s_at	DGKZ	8525	diacylglycerol kinase, zeta 104kDa	3/5	1.0	36829	upregulated	0
201626_at	INSIG1	3638	insulin induced gene 1	3/5	-1.0	42966	downregulated	0
202911_at	MSH6	2956	mutS homolog 6 (E. coli)	3/5	1.0	43110	upregulated	0
201477_s_at	RRM1	6240	ribonucleotide reductase M1	3/5	1.0	43215	upregulated	0
203075_at	SMAD2	4087	SMAD family member 2	3/5	-1.0	46640	downregulated	0
207405_s_at	RAD17	5884	RAD17 homolog (S. pombe)	3/5	1.0	47733	upregulated	0

Table 2: Genes which are specifically upregulated in airway brushings of COPD patients compared with control airway brushings

probeset_ids	Gene Symbol	Entrez Gene Id	Gene Title	present / total lists	fold change in training (COPD vs control)	Rank by abs tvalue	direction of change in COPD vs. control	fc > 2
201587_s_at	IRAK1	3654	interleukin-1 receptor-associated kinase 1	4/5	1.2	2166	upregulated	0
214088_s_at	FUT3	2525	fucosyltransferase 3 (galactoside 3(4)-L-fucosyltransferase, Lewis blood group)	4/5	1.4	11214	upregulated	0
33323_r_at	SFN	2810	stratifin	4/5	1.4	11226	upregulated	0
200931_s_at	VCL	7414	vinculin	4/5	1.2	20026	upregulated	0
203385_at	DGKA	1606	diacylglycerol kinase, alpha 80kDa	4/5	1.1	29676	upregulated	0
202314_at	CYP51A1	1595	cytochrome P450, family 51, subfamily A, polypeptide 1	4/5	1.0	54179	upregulated	0
231126_at	C2orf70	339778	chromosome 2 open reading frame 70	3/5	2.1	3	upregulated	1
206595_at	CST6	1474	cystatin E/M	3/5	2.9	50	upregulated	1
218130_at	C17orf62	79415	chromosome 17 open reading frame 62	3/5	1.4	52	upregulated	0
201761_at	MTHFD2	10797	methylentetrahydrofolate dehydrogenase (NADP+ dependent) 2, methenyltetrahydrofolate cyclohydrolase	3/5	2.0	190	upregulated	1
210845_s_at	PLAUR	5329	plasminogen activator, urokinase receptor	3/5	2.6	467	upregulated	1
203221_at	TLE1	7088	transducin-like enhancer of split 1 (E(spl) homolog, Drosophila)	3/5	1.5	527	upregulated	0
210397_at	DEFB1	1672	defensin, beta 1	3/5	2.6	591	upregulated	1

probeset_ids	Gene Symbol	Entrez Gene Id	Gene Title	present / total lists	fold change in training (COPD vs control)	Rank by absolute value	direction of change in COPD vs. control	fc > 2
204106_at	TESK1	7016	testis-specific kinase 1	3/5	1.3	846	upregulated	0
201666_at	TIMP1	7076	TIMP metalloproteinase inhibitor 1	3/5	1.5	1082	upregulated	0
209906_at	C3AR1	719	complement component 3a receptor 1	3/5	1.9	1568	upregulated	0
227194_at	FAM3B	54097	family with sequence similarity 3, member B	3/5	1.8	1874	upregulated	0
201884_at	CEACAM5	1048	carcinoembryonic antigen-related cell adhesion molecule 5	3/5	3.1	1971	upregulated	1
209369_at	ANXA3	306	annexin A3	3/5	1.5	2088	upregulated	0
205513_at	TCN1	6947	transcobalamin I (vitamin B12 binding protein, R binder family)	3/5	3.4	2605	upregulated	1
210058_at	MAPK13	5603	mitogen-activated protein kinase 13	3/5	1.3	5160	upregulated	0
201141_at	GPNNMB	10457	glycoprotein (transmembrane) nmb	3/5	2.1	5266	upregulated	1
214070_s_at	ATP10B	23120	ATPase, class V, type 10B	3/5	1.5	8213	upregulated	0
205019_s_at	VIPR1	7433	vasoactive intestinal peptide receptor 1	3/5	1.3	9695	upregulated	0
210145_at	PLA2G4A	5321	phospholipase A2, group IVA (cytosolic, calcium-dependent)	3/5	1.4	10939	upregulated	0
205349_at	GNAI5	2769	guanine nucleotide binding protein (G protein), alpha 15 (Gq class)	3/5	1.2	11114	upregulated	0
203655_at	XRCC1	7515	X-ray repair complementing defective repair in Chinese hamster cells 1	3/5	1.1	12911	upregulated	0
201860_s_at	PLAT	5327	plasminogen activator, tissue	3/5	1.4	18596	upregulated	0

probeset_ids	Gene Symbol	Entrez Gene Id	Gene Title	present / total lists	fold change in training (COPD vs control)	Rank by absolute value	direction of change in COPD vs. control	fc > 2
207365_x_at	USP34	9736	ubiquitin specific peptidase 34	3/5	1.1	21072	upregulated	0
204971_at	CSTA	1475	cystatin A (stefin A)	3/5	1.3	21314	upregulated	0
203652_at	MAP3K11	4296	mitogen-activated protein kinase kinase 11	3/5	1.2	22650	upregulated	0
203060_s_at	PAPSS2	9060	3'-phosphoadenosine 5'-phosphosulfate synthase 2	3/5	1.2	25286	upregulated	0
224370_s_at	CAPS2	84698	calcyphosine 2	3/5	1.1	34524	upregulated	0
207556_s_at	DGKZ	8525	diacylglycerol kinase, zeta 104kDa	3/5	1.0	36829	upregulated	0
202911_at	MSH6	2956	mutS homolog 6 (E. coli)	3/5	1.0	43110	upregulated	0
201477_s_at	RRM1	6240	ribonucleotide reductase M1	3/5	1.0	43215	upregulated	0
207405_s_at	RAD17	5884	RAD17 homolog (S. pombe)	3/5	1.0	47733	upregulated	0

Table 3: Genes which are specifically downregulated in airway brushings of COPD patients compared with control airway brushings

probeset_ids	Gene Symbol	Entrez Gene Id	Gene Title	present / total lists	fold change in training (COPD vs control)	Rank by absolute value	direction of change in COPD vs. control	fc > 2
207808_s_at	PROS1	5627	protein S (alpha)	4/5	-2.3	64	downregulated	1
218807_at	VAV3	10451	vav 3 guanine nucleotide exchange factor	4/5	-1.3	2636	downregulated	0
226148_at	ZBTB44	29068	zinc finger and BTB domain containing 44	4/5	-1.1	12186	downregulated	0
214598_at	CLDN8	9073	claudin 8	4/5	-1.4	14469	downregulated	0
213578_at	BMPRI1A	657	bone morphogenetic protein receptor, type 1A	4/5	-1.1	16909	downregulated	0
225761_at	PAPD4	167153	PAP associated domain containing 4	4/5	-1.1	17004	downregulated	0
201877_s_at	PPP2R5C	5527	protein phosphatase 2, regulatory subunit B', gamma	4/5	-1.1	25186	downregulated	0
1556012_at	KLHDC7A	127707	kelch domain containing 7A	3/5	-1.9	87	downregulated	0
1552396_at	WFDC6	140870	WAP four-disulfide core domain 6	3/5	-2.5	101	downregulated	1
221045_s_at	PER3	8863	period homolog 3 (Drosophila)	3/5	-1.7	242	downregulated	0
204369_at	PIK3CA	5290	phosphoinositide-3-kinase, catalytic, alpha polypeptide	3/5	-1.6	446	downregulated	0
211596_s_at	LRIG1	26018	leucine-rich repeats and immunoglobulin-like domains 1	3/5	-1.3	484	downregulated	0
201964_at	SETX	23064	senataxin	3/5	-1.4	533	downregulated	0
222791_at	RSBN1	54665	round spermatid basic protein 1	3/5	-1.5	605	downregulated	0
202816_s_at	SS18	6760	synovial sarcoma translocation, chromosome 18	3/5	-1.4	733	downregulated	0

probeset_ids	Gene Symbol	Entrez Gene Id	Gene Title	present / total lists	fold change in training (COPD vs control)	Rank by absolute value	direction of change in COPD vs. control	fc > 2
207177_at	PTGFR	5737	prostaglandin F receptor (FP)	3/5	-1.7	891	downregulated	0
209189_at	FOS	2353	FBF murine osteosarcoma viral oncogene homolog	3/5	-4.1	1142	downregulated	1
204062_s_at	ULK2	9706	unc-51-like kinase 2 (C. elegans)	3/5	-1.4	1285	downregulated	0
223805_at	OSBPL6	114880	oxysterol binding protein-like 6	3/5	-1.5	1976	downregulated	0
222870_s_at	B3GNT2	10678	UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 2	3/5	-1.2	2429	downregulated	0
201041_s_at	DUSP1	1843	dual specificity phosphatase 1	3/5	-1.7	2469	downregulated	0
203510_at	MET	4233	met proto-oncogene (hepatocyte growth factor receptor)	3/5	-1.3	2537	downregulated	0
209022_at	STAG2	10735	stromal antigen 2	3/5	-1.2	4361	downregulated	0
201941_at	CPD	1362	carboxypeptidase D	3/5	-1.2	4845	downregulated	0
202539_s_at	HMGCR	3156	3-hydroxy-3-methylglutaryl-CoA reductase	3/5	-1.2	6340	downregulated	0
226312_at	RICTOR	253260	RPTOR independent companion of MTOR, complex 2	3/5	-1.2	8387	downregulated	0
203182_s_at	SRPK2	6733	SRSF protein kinase 2	3/5	-1.1	10114	downregulated	0
213070_at	PIK3C2A	5286	phosphoinositide-3-kinase, class 2, alpha polypeptide	3/5	-1.2	11271	downregulated	0
209218_at	SQLE	6713	squalene epoxidase	3/5	-1.2	12374	downregulated	0
1558486_at	ZNF493	284443	zinc finger protein 493	3/5	-1.3	12655	downregulated	0
200985_s_at	CD59	966	CD59 molecule, complement regulatory protein	3/5	-1.1	13226	downregulated	0
208881_x_at	ID11	3422	isopentenyl-diphosphate delta isomerase 1	3/5	-1.2	13569	downregulated	0

probeset_ids	Gene Symbol	Entrez Gene Id	Gene Title	present / total lists	fold change in training (COPD vs control)	Rank by absolute value	direction of change in COPD vs. control	fc > 2
205080_at	RARB	5915	retinoic acid receptor, beta	3/5	-1.2	13658	downregulated	0
212533_at	WEE1	7465	WEE1 homolog (S. pombe)	3/5	-1.1	14607	downregulated	0
219485_s_at	PSMD10	5716	proteasome (prosome, macropain) 26S subunit, non-ATPase, 10	3/5	-1.1	14730	downregulated	0
208820_at	PTK2	5747	PTK2 protein tyrosine kinase 2	3/5	-1.1	15909	downregulated	0
201384_s_at	NBR1	4077	Neighbor of BRCA1 gene 1	3/5	-1.1	17242	downregulated	0
227038_at	SGMS2	166929	sphingomyelin synthase 2	3/5	-1.1	20534	downregulated	0
225240_s_at	MSI2	124540	musashi homolog 2 (Drosophila)	3/5	-1.1	20773	downregulated	0
202277_at	SPTLC1	10558	serine palmitoyltransferase, long chain base subunit 1	3/5	-1.1	22925	downregulated	0
229738_at	DNAH10	196385	dynein, axonemal, heavy chain 10	3/5	-1.1	24994	downregulated	0
202603_at	ADAM10	102	ADAM metalloproteinase domain 10	3/5	-1.1	27094	downregulated	0
201528_at	RPA1	6117	replication protein A1, 70kDa	3/5	-1.1	28648	downregulated	0
214484_s_at	SIGMAR1	10280	sigma non-opioid intracellular receptor 1	3/5	-1.1	31396	downregulated	0
227520_at	TXLNG	55787	taxilin gamma	3/5	-1.1	32441	downregulated	0
201626_at	INSIG1	3638	insulin induced gene 1	3/5	-1.0	42966	downregulated	0
203075_at	SMAD2	4087	SMAD family member 2	3/5	-1.0	46640	downregulated	0

5 We claim:

1. A method of diagnosing an individual as being at risk for or having Chronic Obstructive Pulmonary Disease (COPD) comprising detecting the expression level of at least 2 of the 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
10 genes in the test sample diagnoses the individual as suffering from COPD.
2. The method according to claim 1, 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.
- 15 3. The method according to claim 1 or 2, 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
20 vector machine, dual bagging, single and pairs, forward learning, Laplacian based 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 GSE10106, GSE10135, GSE11906, GSE11952, GSE13933, GSE19407,
25 GSE19667, GSE20257, GSE5058, GSE7832, and GSE8545.
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
30 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: PROS1, IRAK1, VAV3, FUT3,

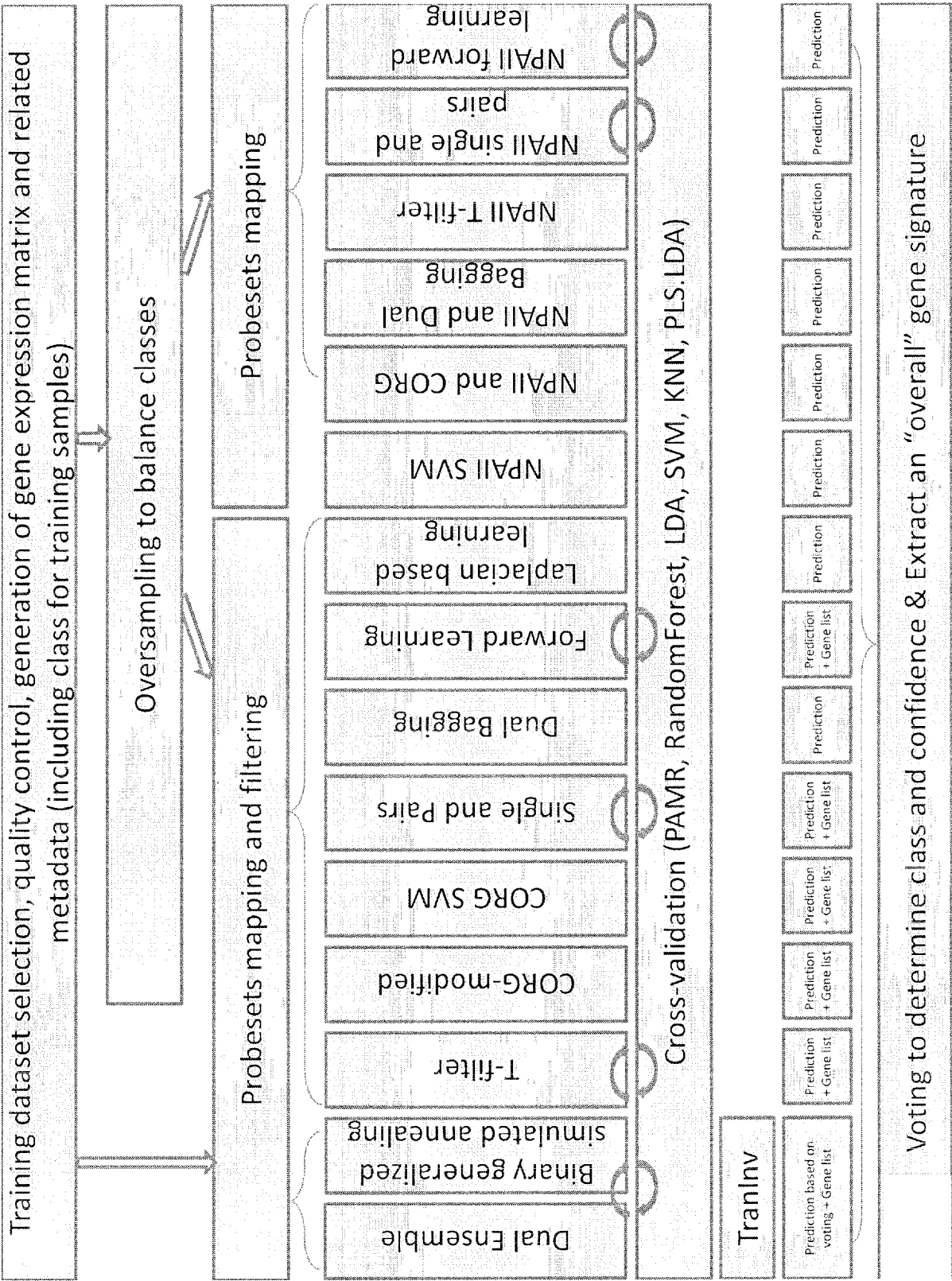
SFN, ZBTB44, CLDN8, BMPR1A, PAPD4, VCL, PPP2R5C, DGKA, and CYP51A1.

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.
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.
9. An array comprising polynucleotides hybridizing to at least 2 COPD signature genes immobilized on a solid surface, wherein the COPD signature genes are selected from the genes listed in Table 1 and said array is not a human genome-wide array.
10. A device comprising antibodies immobilized on a solid surface that bind to proteins encoded by at least 2 COPD signature genes, wherein the COPD signature genes are selected from the genes listed in Table 1.
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.
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 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 GSE10106, GSE10135, GSE11906, GSE11952, GSE13933, GSE19407, GSE19667, GSE20257, GSE5058, GSE7832, and GSE8545.

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 PROS1, IRAK1, VAV3, FUT3, SFN, ZBTB44, CLDN8, BMPR1A, PAPD4, VCL, PPP2R5C, DGKA, and CYP51A1.

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

Figure 1



INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2013/062996

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, CHEM ABS Data, EMBASE, FSTA, 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	GEORGE L MUTTER ET AL: "Global Expression Changes of Constitutive and Hormonally Regulated Genes During Endometrial Neoplastic Transformation", GYNECOLOGIC ONCOLOGY, ACADEMIC PRESS, LONDON, GB, vol. 83, no. 2, 1 November 2001 (2001-11-01), pages 177-185, XP008110295, ISSN: 0090-8258, DOI: 10.1006/GYNO.2001.6352 [retrieved on 2002-03-04] the whole document ----- -/--	9,15



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

3 September 2013

Date of mailing of the international search report

12/09/2013

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

Leber, Thomas

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2013/062996

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	"Shows that not genome-wide array HU6800 comprises probe for Fut3. The array is disclosed in XP008110295", 1 January 2001 (2001-01-01), XP55076507, Retrieved from the Internet: URL:http://www.affymetrix.com [retrieved on 2013-08-26] the whole document	9,15
X	"Shows that not genome-wide array HU6800 comprises probe for IRAK1. The array is disclosed in XP008110295", 1 January 2001 (2001-01-01), XP055076505, Retrieved from the Internet: URL:http://www.affymetrix.com [retrieved on 2013-08-26] the whole document	9,15
X	"Shows that not genome-wide array HU6800 comprises probe for PROS1. The array is disclosed in XP008110295", 1 January 2001 (2001-01-01), XP055076503, Retrieved from the Internet: URL:http://www.affymetrix.com [retrieved on 2013-08-26] the whole document	9,15
X	ELISA PENNA ET AL: "microRNA-214 contributes to melanoma tumour progression through suppression of TFAP2C", THE EMBO JOURNAL, vol. 30, no. 10, 18 May 2011 (2011-05-18), pages 1990-2007, XP055076598, ISSN: 0261-4189, DOI: 10.1038/emboj.2011.102 page 2005, right col.	10,15
X	"Product Insert of Proteome Profiler Array "Human Soluble Receptor Array Kit" antibody array described in XP055076598", 1 February 2012 (2012-02-01), XP055076601, Retrieved from the Internet: URL:http://www.rndsystems.com/pdf/ary012.pdf [retrieved on 2013-08-27] page 3, table page 15, "Coordinate A11, A12"; page 16, "Coordinate F3, F4" -/--	10,15

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2013/062996

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2006/105252 A2 (UNIV COLORADO [US]; GERACI MARK W [US]; COLDREN CHRISTOPHER D [US]; GR) 5 October 2006 (2006-10-05)	11-14
Y	claim 1; page 66 "Microarray data analysis"	1-8
Y	WO 2008/109773 A2 (CORNELL RES FOUNDATION INC [US]; CRYSTAL RONALD G [US]; HACKETT NEIL R) 12 September 2008 (2008-09-12) claims 1 and 3	1-8
Y	LIU L ET AL: "Detection of circulating cancer cells in lung cancer patients with a panel of marker genes", BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS, ACADEMIC PRESS INC. ORLANDO, FL, US, vol. 372, no. 4, 8 August 2008 (2008-08-08), pages 756-760, XP027181572, ISSN: 0006-291X, DOI: 10.1016/J.BBRC.2008.05.101 [retrieved on 2008-06-02] page 757, right col.; Fig. 2	1-8
Y	PANICI P ET AL: "Predictive value of multiple tumor marker assays in second-look procedures for ovarian cancer", GYNECOLOGIC ONCOLOGY, ACADEMIC PRESS, LONDON, GB, vol. 35, no. 3, 1 December 1989 (1989-12-01), pages 286-289, XP026231880, ISSN: 0090-8258, DOI: 10.1016/0090-8258(89)90064-4 [retrieved on 1989-12-01] abstract	1-8

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/EP2013/062996

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2006105252	A2	05-10-2006	NONE

WO 2008109773	A2	12-09-2008	US 2010119474 A1 13-05-2010
		WO 2008109773 A2	12-09-2008
