(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
20 September 2012 (20.09.2012)

WIPO | PCT

(10) International Publication Number
WO 2012/125712 A2

(51) International Patent Classification:
*C12Q 1/68* (2006.01) *C12N 15/11* (2006.01)
*G01N 33/68* (2006.01)

(21) International Application Number:
PCT/US2012/029056

(22) International Filing Date:
14 March 2012 (14.03.2012)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/452,481 14 March 2011 (14.03.2011) US

(71) Applicant *(for all designated States except US)*: **RESPIRA HEALTH, LLC** [US/US]; 667 Madison Avenue, New York, NY 10065 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **ALBINO, Anthony** [US/US]; 215 West 92nd St., New York, NY 10025 (US). **HERNANDEZ, Joseph** [US/US]; 6053 White Flint Drive, Frederick, MD 21702 (US). **VAN LAAR, Ryan** [AU/US]; 208 East 88th Street, 2D, New York, NY 10128 (US).

(74) Agents: **TANNOCH-MAGIN, Vivien, J.** et al.; Hamilton, Brook, Smith & Reynolds, P.C., 530 Virginia Rd., P.O. Box 9133, Concord, MA 01742-9133 (US).

(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, 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, KM, 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, 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, MD, 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, ML, MR, NE, SN, TD, TG).

**Published**:

— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*

WO 2012/125712 A2

(54) Title: LUNG TUMOR CLASSIFIER FOR CURRENT AND FORMER SMOKERS

(57) Abstract: The invention provides methods for the diagnosis of lung cancer in a subject based detecting abnormal gene expression patterns. The invention further provides methods of treating lung cancer.

# LUNG TUMOR CLASSIFIER FOR CURRENT AND FORMER SMOKERS

## RELATED APPLICATION(S)

This application claims the benefit of U.S. Provisional Application No. 61/452,481, filed on March 14, 2011.

The entire teachings of the above application are incorporated herein by reference.

## BACKGROUND OF THE INVENTION

Lung cancer has an extremely high mortality rate of up to 85% within five years and is consequently the number one cause of cancer death in the US and the rest of the world. Current or former smokers can often be suspect for lung cancer on the basis of unusual radiographic imaging results or the presence of clinical symptoms associated with the disease. The technique of flexible bronchoscopy is used as a noninvasive diagnostic test in these cases; however the sensitivity for identifying individuals who actually have lung cancer can vary from 30-80%.

Thus, there is a need for better tests to diagnose lung cancer in current and former smokers.

## SUMMARY OF THE INVENTION

The present invention provides, *inter alia*, diagnostic methods for lung cancer and associated methods of treatment. The invention is based, at least in part, on the discovery that a support vector machine (SVM) index, based on the expression level of 51 genes, is a significant diagnostic tool for lung cancer.

Significant diagnostic power can be achieved with smaller subsets of genes, such as three of the 51.

Accordingly, in one aspect, the invention provides a method of predicting the risk of lung cancer in a subject. The method comprises the step of determining (*e.g.*, by testing by any means), in an isolated sample from a patient (*e.g.*, a biological sample), whether the sample exhibits an abnormal expression pattern of one or more genes selected from ABCF2, AGMAT, BBS9, BLK, C16orf42, CAT, CDK14, CLCN3, CNPY4, DIP2A, ENPP4, EPX, EXOC6B, EZH1, H2AFV, HAUS2, IL1F6, INHBC, IVD, MCF2L, MRPL52, MYL12B, MZT2B, OLFML2B, ORC6L, PDE4C, PGF, PGRMC1, POLR1B, PRDX2, PRKAR1A, RIOK3, RNASE4, RPL27A, RPL37A, RPL38, RRAGD, SBF1, SH2B2, SLC35E1, SLC39A6, SNX17, SPG21, TAOK1, TEX13B, TYMP, USP34, WDR6, XRCC2, ZNF701, and ZNF721, where an abnormal expression pattern predicts and increased risk of lung cancer in the patient. In particular embodiments, the method comprises determining the expression level of at least three

of these genes, *e.g.*, at least 3, 5, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 35, 40, 45, 50, or all 51 genes. In still more particular embodiments, the expression level of at least 3 genes is determined. In other particular embodiments, the expression level of at least 7 genes is determined.

In some embodiments, at least one of the genes is selected from ABCF2, AGMAT, BBS9, BLK, C16orf42, CAT, CDK14, CNPY4, DIP2A, ENPP4, EPX, EXOC6B, EZH1, H2AFV, HAUS2, IL1F6, IVD, MCF2L, MYL12B, MZT2B, OLFML2B, ORC6L, PDE4C, PGRMC1, RPL27A, RRAGD, SBF1, SH2B2, SNX17, SPG21, TAOK1, TEX13B, TYMP, WDR6, XRCC2, ZNF701, and ZNF721.

In certain embodiments, at least one of the genes is selected from CAT, ENPP4, RRAGD, MCF2L, SLC39A6, SNX17, PGRMC1, ZNF721, PRKAR1A, CLCN3, RNASE4, BBS9, H2AFV, POLR1B, DIP2A, EZH1, MRPL52, HAUS2, RIOK3, MZT2B, *e.g.*, the expression level of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, or all 20 genes is determined. In more particular embodiments, at least one of the genes is selected from CAT, ENPP4, RRAGD, MCF2L, SNX17, PGRMC1, ZNF721, BBS9, H2AFV, DIP2A, EZH1, HAUS2, and MZT2B, *e.g.*, the expression level of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, or all 13 genes is determined. In still more particular embodiments, the expression levels of CAT, ENPP4, and RRAGD are determined.

In some embodiments an abnormal expression pattern comprises increased expression of one or more of BBS9, CAT, CLCN3, ENPP4, EZH1, H2AFV, MCF2L, PGRMC1, PRKAR1A, RNASE4, RRAGD, SLC39A6, or SNX17. In certain embodiments, an abnormal expression pattern comprises decreased expression of one or more of ABCF2, AGMAT, BLK, C16orf42, CDK14, CNPY4, DIP2A, EPX, EXOC6B, FAM128B, HAUS2, IL1F6, INHBC, IVD, MRPL52, MYL12B, OLFML2B, ORC6L, PDE4C, PGF, POLR1B, PRDX2, RIOK3, RPL27A, RPL37A, RPL38, SBF1, SH2B2, SLC35E1, SPG21, TAOK1, TEX13B, TYMP, USP34, WDR6, XRCC2, ZNF701 or ZNF721. In more particular embodiments, an abnormal expression pattern comprises increased expression of one or more of BBS9, CAT, CLCN3, ENPP4, EZH1, H2AFV, MCF2L, PGRMC1, PRKAR1A, RNASE4, RRAGD, SLC39A6, or SNX17 and decreased expression of one or more of ABCF2, AGMAT, BLK, C16orf42, CDK14, CNPY4, DIP2A, EPX, EXOC6B, FAM128B, HAUS2, IL1F6, INHBC, IVD, MRPL52, MYL12B, OLFML2B, ORC6L, PDE4C, PGF, POLR1B, PRDX2, RIOK3, RPL27A, RPL37A, RPL38,

SBF1, SH2B2, SLC35E1, SPG21, TAOK1, TEX13B, TYMP, USP34, WDR6, XRCC2, ZNF701 or ZNF721.

The expression level of a gene may be measured at the nucleic acid (*e.g.*, mRNA) or protein level. In particular embodiments, the expression level of one or more genes is measured at the nucleic acid level and in more particular embodiments, the expression levels of two or more genes are measure simultaneously, for example, using a microarray. In still more particular embodiments, the expression level of one or more genes is measured at the nucleic acid level using a microarray, such as an Exon 1.0 ST, Gene 1.0 ST, U 95, U133, U133A 2.0, or U133 Plus 2.0 AFYYMETRIX™ microarray. In other embodiments, the expression levels of one or more genes is measured at the nucleic acid level by rtPCR followed by qPCR, or serial analysis of gene expression (SAGE).

Expression levels can be analyzed by any means known in the art. In certain embodiments, for example, expression levels can transformed before evaluation, *e.g.*, expressed as a fold-induction, log normalized, and, optionally, percentile ranked. In particular embodiments, a sample is classified using by an SVM. The SVM may be based on the expression level of any of the gene combinations described above. In still more particular embodiments, the SVM may use weights substantially similar (*i.e.*, within about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 40, or 50%) to MCF2L (-1.4179), CLCN3 (-1.3756), EZH1 (-1.3011), TYMP (1.0144), CDK14 (0.8952), IVD (0.8668), EXOC6B (0.8071), ABCF2 (0.7865), SBF1 (0.7841), RRAGD (-0.7208), MRPL52 (0.7176), CAT (-0.6615), INHBC (0.6389), RIOK3 (0.6296), C16orf42 (0.5584), TEX13B (0.5532), RPL37A (-0.5478), SNX17 (-0.4909), AGMAT (0.4859), BBS9 (-0.4511), WDR6 (0.4187), SPG21 (0.4007), MZT2B (-0.3857), RPL38 (-0.3581), OLFML2B (-0.3557), H2AFV (0.348), ENPP4 (-0.3385), ZNF721 (0.2899), PRDX2 (0.2413), MYL12B (0.2217), RPL27A (-0.2036), ZNF701 (0.1925), SLC39A6 (-0.1778), EPX (-0.1525), PDE4C (-0.1521), DIP2A (0.152), IL1F6 (0.1517), POLR1B (0.1501), RNASE4 (0.1278), ORC6L (-0.1093), XRCC2 (-0.1071), PRKAR1A (-0.0991), PGF (-0.0949), USP34 (-0.0769), PGRMC1 (-0.0763), CNPY4 (0.0726), SLC35E1 (0.0713), TAOK1 (0.0602), SH2B2 (0.0529), BLK (-0.0373), and HAUS2 (-0.0301).

A patient to be tested by the methods provided by the invention may be symptomatic or asymptomatic for lung cancer. In certain embodiments, the patient is a current or former smoker. In more particular embodiments, the patient is a former smoker. In still more particular embodiments, the patient is a former smoker who quit smoking within 10 years of isolation of

the sample to be tested. In some embodiments, the subject is a non-smoker but may be considered to be at an increased risk for developing lung cancer—for example, the subject may have a history of exposure to second hand tobacco smoke, other environmental exposure, or genetic predisposition.

A sample to be tested for abnormal gene expression can be obtained from a patient by any means known in the art from any suitable source in the oral, esophageal, nasal, and/or pulmonary system. In particular embodiments, the sample is obtained by bronchoscopy or from nasal epithelial tissue.

In certain embodiments, the methods provided by the invention may further comprise the step of follow-on diagnosis including one or more of sputum cytology, flexible bronchoscopy (FB), transthoracic needle aspiration (TTNA), 18F-Fluorodeoxyglucose-positron emission tomography (FDG-PET), magnetic resonance imaging (MRI), endobronchial ultrasound and conventional or low-dose spiral computed tomography (LDCT), chest X-ray, or any biopsy. In more particular embodiments, subsequent to testing by the methods provided by the invention, the subject is determined to have lung cancer. In still more particular embodiments, the subject is determined to have early stage lung cancer, such as stage 1a or 1b.

Accordingly, in another aspect the invention provides a method of treating lung cancer comprising administering a suitable prophylaxis to a patient determined to have lung cancer by any of the methods provided by the invention. In particular embodiments, the prophylaxis includes, *e.g.*, chemotherapy, hormonal therapy, immunotherapy, radiotherapy, surgery, targeted gene therapies (*e.g.*, epidermal growth factor receptor-tyrosine kinase inhibitors, such as gefitinib; and agents targeting ALK mutations and rearrangements, such as crizotinib, etc.) and combinations thereof.

## BRIEF DESCRIPTION OF THE DRAWINGS

The patent or application file contains at least one drawing executed in color. Copies of this patent or patent application publication with color drawing(s) will be provided by the Office upon request and payment of the necessary fee.

FIG. 1 is a representation of a hierarchical clustering of the 51 gene signature. Patients are represented by columns (1 = diagnosed with lung cancer, 0 = cancer-free) and genes are represented by rows.

FIGS. 2a-2e summarize significantly enriched biological function categories in the 51-gene signature identified by Ingenuity Pathway Analysis.

FIG. 3 summarizes the results of a ROC analysis of SVM index for the 90 patient training series.

FIG. 4 summarizes the gene expression data and clinical variables for the 90 patient training series. The yellow horizontal line across the figure represents the classification threshold of -0.16. Patients below this line are classified as 'non-tumor' while patients above the line are classified as 'tumor'. To the left of the gene expression data (red = high relative expression, green = low relative expression, in tumor vs. non-tumor) the patient ID, tumor/non-tumor status and SVM index is shown. To the right of the figure is each patient's bronchoscopy status, age, sex and cigarette pack-years. Patient data is also summarized in Table 5a.

FIG. 5 is a graph summarizing classifier LOOCV performance for SVM algorithms calculated on gene subsets of the 51 gene signature.

FIG. 6 is a graph of the mean of all LOOCV classification performance criteria shown in FIG. 5, generated for variations of the SVM classification algorithm using gene subsets of the 51 gene signature.

FIG. 7 summarizes the 51 gene signatures in the 60 patient validation, series 1. Patients represented by rows and genes are represented by columns. Red = high expression, green = low expression. The yellow horizontal line corresponds to a classification threshold of -0.75. Patient data is also summarized in Table 5b.

FIG. 8 is a graphical summary of a ROC analysis of the 51-gene SVM classifier applied to the 60 patient validation series 1. AUC = 0.78 (95% CI: 0.65 to 0.87) $P<0.0001$.

FIG. 9 is a graphical summary of a ROC analysis of the 51-gene SVM and a previously-published [1] 80-gene algorithm for predictions of validation series 1.

FIG. 10 is a box plot of the 51 gene SVM indices calculated from lung tumor and normal lung tissue gene expression profiles. The difference is statistically significant ($P<=0.001$).

FIG. 11 summarizes a ROC analysis of the 51-gene SVM applied to unpaired specimens from independent validation series 2, consisting of 41 gene expression profiles from either lung tumor or normal lung tissue. The AUC of 0.76 was statistically significant ($P=0.0006$).

FIG. 12 is a box plot of 51-gene SVM indices calculated from gene expression data generated from histologically-normal BEC's from individuals of varying smoking status. $P=0.25$.

FIG. 13 is a general linear model analysis of the 51 gene SVM index measured in BEC of histologically normal, current and former smokers.

FIG. 14 summarizes exemplary steps involved in applying the 51-gene SVM to gene expression data generated from lung BEC and making a prediction of cancer or non-cancer.

## DETAILED DESCRIPTION OF THE INVENTION

The invention provides methods for the diagnosis of lung cancer. The methods provided by the invention may be used to classify a patient as having lung cancer or not having lung cancer based on an abnormal gene expression pattern. The invention further provides identifying patients for treatment of lung cancer. The following definitions are will be adhered to throughout this application.

"Patient" refers to a human at any stage of development. Examples of suitable patients include, but are not limited to, both female and male adult patients that have, or are at risk for developing, a lung cancer.

"Gene expression" refers to both nucleic acid level (*e.g.*, mRNA or cDNA derived from it) and protein level expression of a gene. Genes expressed as nucleic acids may or may not encode for and/or be translated into protein.

"Level of expression," "expression level," "gene expression level" and the like, refers to the amount of a gene expression product (*e.g.*, mRNA or protein). Expression levels may be transformed (*e.g.*, normalized) or analyzed "raw."

"Expression pattern" means at least two expression levels. For example, in some embodiments, the two or more expression levels may be the expression level of one gene at two or more time points or the expression levels of two or more different genes at the same, or different, times.

"Abnormal expression pattern" refers to a significant statistical and/or practical deviation in the expression level of one or more genes, relative to a suitable control. "Suitable controls" include, for example, paired biopsy samples from a single patient (*e.g.*, tissue samples obtained at different times from a patient, *e.g.*, before and after developing cancer; as well as a pair of biopsies from morphologically and/or histologically normal and morphologically and/or histologically suspect samples from the patient, which may be obtained at the same or different times) as well as reference values previously compiled from samples determined—by any means—to be cancerous or non-cancerous. For example, reference values for one or more genes

may be compiled and used to develop a binary or probabilistic classification algorithm that is then used to classify a sample from a patient as cancerous or non-cancerous.

"Highly stringent hybridization" means hybridization conditions comprising about 6X SSC and 1% SDS at 65°C, with a first wash for 10 minutes at about 42°C with about 20% (v/v) formamide in 0.1X SSC, and with a subsequent wash with 0.2 X SSC and 0.1% SDS at 65°C.

*Genes for use in the methods provided by the invention*

The invention provides a useful set of genes that are differentially expressed in cancerous and non-cancerous lung tissues that can be used to diagnose cancer in a variety of patient samples from the oral, esophageal, nasal, and/or pulmonary system. The genes identified by the present invention are listed in Table 1. Table 1 further provides reference GeneIDs, mRNA sequence accession numbers, protein sequence accession numbers, and AffymetrixID. These identifiers may be used to retrieve, *inter alia*, publicly-available annotated mRNA or protein sequences from sources such as the NCBI website, which may be found at the following uniform resource locator (URL): http://www.ncbi.nlm.nih.gov. The information associated with these identifiers, including reference sequences and their associated annotations, are all incorporated by reference. Additional useful tools for converting IDs or obtaining additional information on a gene are known in the art and include, for example, DAVID, Clone/GeneID converter and SNAD. *See* Huang e*t al., Nature Protoc.* 4(1):44-57 (2009), Huang *et al., Nucleic Acids Res.* 37(1)1-13 (2009), Alibes *et al., BMC Bioinformatics* 8:9 (2007), Sidorov *et al., BMC Bioinformatics* 10:251 (2009).

Table 1

| Symbol | Name | Unigene ID | Cytoband | Affymetrix ProbeSetID | Entrez GeneID | RefSeq RNA ID | RefSeq peptide |
|---|---|---|---|---|---|---|---|
| ABCF2 | ATP-binding cassette, sub-family F (GCN20), member 2 | Hs.654958 | 7q36 | 209246_at | 10061 | NM_005692<br>NM_007189 | NP_005683.2<br>NP_009120.1 |
| AGMAT | agmatine ureohydrolase (agmatinase) | Hs.567583 | 1p36.21 | 219792_at | 79814 | NM_024758.3 | NP_079034.3 |
| BBS9 | Bardet-Biedl syndrome 9 | Hs.372360 | 7p14 | 209958_s_at | 27241 | NM_001033604.1<br>NM_001033605<br>NM_014451<br>NM_198428.2 | NP_001028776.1<br>NP_001028777.1<br>NP_055266.2<br>NP_940820.1 |
| BLK | B lymphoid tyrosine kinase | Hs.146591 | 8p23-p22 | 210934_at | 640 | NM_001715<br>NM_001170 | NP_001706.2<br>NP_001161.1 |
| C16orf42 | chromosome 16 open reading frame 42 | Hs.134846 | 16p13.3 | 213105_s_at | 115939 | NM_001001410.2 | NP_001001410.1 |
| CAT | catalase | Hs.502302 | 11p13 | 201432_at | 847 | NM_001752 | NP_001743.1 |
| CDK14 | cyclin-dependent kinase 14 | Hs.430742 | 7q21-q22 | 211502_s_at | 5218 | NM_012395 | NP_036527.1 |
| CLCN3 | chloride channel 3 | Hs.481186 | 4q33 | 201732_s_at | 1182 | NM_001829<br>NM_173872 | NP_001820.2<br>NP_776297.2 |

| Symbol | Name | Unigene ID | Cytoband | Affymetrix ProbeSetID | Entrez GeneID | RefSeq RNA ID | RefSeq peptide |
|---|---|---|---|---|---|---|---|
| CNPY4 | canopy 4 homolog (zebrafish) | Hs.632293 | 7q22.1 | 217700_at | 245812 | NM_152755 | NP_689968.1 |
| DIP2A | DIP2 disco-interacting protein 2 homolog A (Drosophila) | Hs.189585 | 21q22.3 | 215529_x_at | 23181 | NM_015151.2<br>NM_206889.1<br>NM_206890.1<br>NM_206891.1 | NP_055966.2<br>NP_996772.1<br>NP_996773.1<br>NP_996774.1 |
| ENPP4 | ectonucleotide pyrophosphatase/phosphodiesterase 4 (putative) | Hs.643497 | 6p21.1 | 204160_s_at | 22875 | NM_014936 | NP_055751.1 |
| EPX | eosinophil peroxidase | Hs.279259 | 17q23.1 | 214627_at | 8288 | NM_000502.4 | NP_000493.1 |
| EXOC6B | exocyst complex component 6B | Hs.303454 | 2p13.2 | 215417_at | 23233 | NM_015189.1 | NP_056004.1 |
| EZH1 | enhancer of zeste homolog 1 (Drosophila) | Hs.194669 | 17q21.1-q21.3 | 203249_at | 2145 | NM_001991.3 | NP_001982.2 |
| H2AFV | H2A histone family, member V | Hs.488189 | 7p13 | 212206_s_at | 94239 | NM_012412<br>NM_138635<br>NM_201436<br>NM_201516.2<br>NM_201517 | NP_036544.1<br>NP_619541.1<br>NP_958844.1<br>NP_958924.1<br>NP_958925.1 |
| HAUS2 | HAUS augmin-like complex, subunit 2 | Hs.14347 | 15q15.2 | 220071_x_at | 55142 | NM_018097<br>NM_001130447.1 | NP_060567.1<br>NP_001123919.1 |

| Symbol | Name | Unigene ID | Cytoband | Affymetrix ProbeSetID | Entrez GeneID | RefSeq RNA ID | RefSeq peptide |
|---|---|---|---|---|---|---|---|
| IL1F6 | interleukin 1 family, member 6 (epsilon) | Hs.278910 | 2q12-q14.1 | 221404_at | 27179 | NM_014440.1 | NP_055255.1 |
| INHBC | inhibin, beta C | Hs.632722 | 12q13.1 | 207688_s_at | 3626 | NM_005538.2 | NP_005529.1 |
| IVD | isovaleryl-CoA dehydrogenase | Hs.513646 | 15q14-q15 | 216495_x_at | 3712 | NM_004116<br>NM_054033 | NP_004107.1<br>NP_473374.1 |
| MCF2L | MCF.2 cell line derived transforming sequence-like | Hs.170422 | 13q34 | 212935_at | 23263 | NM_024979 | NP_079255.2 |
| MRPL52 | mitochondrial ribosomal protein L52 | Hs.355935 | 14q11.2 | 221997_s_at | 122704 | NM_178336.1<br>NM_180982<br>NM_181304.1<br>NM_181305.1<br>NM_181306.1<br>NM_181307 | NP_848026.1<br>NP_851313.1<br>NP_851821.1<br>NP_851822.1<br>NP_851823.1<br>NP_851824.1 |
| MYL12B | myosin, light chain 12B, regulatory | Hs.190086 | 18p11.31 | 221474_at | 103910 | NM_001144944.1<br>NM_001144945.1<br>NM_033546 | NP_001138416.1<br>NP_001138417.1<br>NP_291024.1 |
| MZT2B | mitotic spindle organizing protein 2B | Hs.469925 | 2q21.1 | 220720_x_at | 80097 | NM_025029.3 | NP_079305.2 |
| OLFML2B | olfactomedin-like 2B | Hs.507515 | 1q23.3 | 213125_at | 25903 | NM_015441 | NP_056256.1 |

| Symbol | Name | Unigene ID | Cytoband | Affymetrix ProbeSetID | Entrez GeneID | RefSeq RNA ID | RefSeq peptide |
|---|---|---|---|---|---|---|---|
| ORC6L | origin recognition complex, subunit 6 like (yeast) | Hs.49760 | 16q12 | 219105_x_at | 23594 | NM_014321 | NP_055136.1 |
| PDE4C | phosphodiesterase 4C, cAMP-specific | Hs.132584 | 19p13.11 | 206792_x_at | 5143 | NM_000923<br>NM_001098818.1<br>NM_001098819.1 | NP_000914.2<br>NP_001092288.1<br>NP_001092289.1 |
| PGF | placental growth factor | Hs.252820 | 14q24.3 | 215179_x_at | 5228 | NM_001039550.1<br>NM_006736 | NP_001034639.1<br>NP_006727.2 |
| PGRMC1 | progesterone receptor membrane component 1 | Hs.90061 | Xq22-q24 | 201120_s_at | 10857 | NM_006667 | NP_006658.1 |
| POLR1B | polymerase (RNA) I polypeptide B, 128kDa | Hs.86337 | 2q13 | 220113_x_at | 84172 | NM_019014 | NP_061887.2 |
| PRDX2 | peroxiredoxin 2 | Hs.432121 | 19p13.2 | 215067_x_at | 7001 | NM_005809.4<br>NM_181738.1 | NP_005800.3<br>NP_859428.1 |
| PRKAR1A | protein kinase, cAMP-dependent, regulatory, type I, alpha (tissue specific extinguisher 1) | Hs.280342 | 17q23-q24 | 200605_s_at | 5573 | NM_002734<br>NM_212471<br>NM_212472 | NP_002725.1<br>NP_997636.1<br>NP_997637.1 |
| RIOK3 | RIO kinase 3 (yeast) | Hs.445511 | 18q11.2 | 215588_x_at | 8780 | NM_003831 | NP_003822.2 |

| Symbol | Name | Unigene ID | Cytoband | Affymetrix ProbeSetID | Entrez GeneID | RefSeq RNA ID | RefSeq peptide |
|---|---|---|---|---|---|---|---|
| RNASE4 | ribonuclease, RNase A family, 4 | Hs.283749 | 14q11.1 | 205158_at | 6038 | NM_002937<br>NM_194431 | NP_002928.1<br>NP_919412.1 |
| RPL27A | ribosomal protein L27a | Hs.523463 | 11p15 | 212044_s_at | 6157 | NM_000990 | NP_000981.1 |
| RPL37A | ribosomal protein L37a | Hs.433701 | 2q35 | 214041_x_at | 6168 | NM_000998 | NP_000989.1 |
| RPL38 | ribosomal protein L38 | Hs.380953 | 17q25.1 | 221943_x_at | 6169 | NM_000999<br>NM_001035258 | NP_000990.1<br>NP_001030335.1 |
| RRAGD | Ras-related GTP binding D | Hs.31712 | 6q15-q16 | 221524_s_at | 58528 | NM_021244 | NP_067067.1 |
| SBF1 | SET binding factor 1 | Hs.589924 | 22q13.33 | 213383_at | 6305 | NM_002972.2 | NP_002963.2 |
| SH2B2 | SH2B adaptor protein 2 | Hs.489448 | 7q22 | 205367_at | 10603 | NM_020979.2 | NP_066189.2 |
| SLC35E1 | solute carrier family 35, member E1 | Hs.620596 | 19p13.11 | 220796_x_at | 79939 | NM_024881.4 | NP_079157.3 |
| SLC39A6 | solute carrier family 39 (zinc transporter), member 6 | Hs.725276 | 18q12.2 | 202089_s_at | 25800 | NM_001099406.1<br>NM_012319.3 | NP_001092876.1<br>NP_036451.3 |
| SNX17 | sorting nexin 17 | Hs.278569 | 2p23-p22 | 200991_s_at | 9784 | NM_014748 | NP_055563.1 |
| SPG21 | spastic paraplegia 21 (autosomal recessive, Mast syndrome) | Hs.242458 | 15q21-q22 | 215383_x_at | 51324 | NM_016630 | NP_057714.1 |
| TAOK1 | TAO kinase 1 | Hs.631758 | 17q11.2 | 216310_at | 57551 | NM_020791 | NP_065842.1 |
| TEX13B | testis expressed 13B | Hs.333130 | Xq22.3 | 221034_s_at | 56156 | NM_031273 | NP_112563.1 |

| Symbol | Name | Unigene ID | Cytoband | Affymetrix ProbeSetID | Entrez GeneID | RefSeq RNA ID | RefSeq peptide |
|---|---|---|---|---|---|---|---|
| TYMP | thymidine phosphorylase | Hs.592212 | 22q13 | 217497_at | 1890 | NM_001113755.1<br>NM_001113756.1<br>NM_001953.3 | NP_001107227.1<br>NP_001107228.1<br>NP_001944.1 |
| USP34 | ubiquitin specific peptidase 34 | Hs.644708 | 2p15 | 207365_x_at | 9736 | NM_014709.3 | NP_055524.3 |
| WDR6 | WD repeat domain 6 | Hs.654815 | 3p21.31 | 217734_s_at | 11180 | NM_018031.3 | NP_060501.3 |
| XRCC2 | X-ray repair complementing defective repair in Chinese hamster cells 2 | Hs.647093 | 7q36.1 | 207598_x_at | 7516 | NM_005431 | NP_005422.1 |
| ZNF701 | zinc finger protein 701 | Hs.235167 | 19q13.41 | 220242_x_at | 55762 | NM_018260<br>NM_001172655.1 | NP_060730.1<br>NP_001166126.1 |
| ZNF721 | zinc finger protein 721 | Hs.428360 | 4p16.3 | 215978_x_at | 170960 | NM_133474.2 | NP_597731.2 |

An expression pattern for a patient can be obtained by determining the expression level of one or more of the genes in Table 1. In particular embodiments, the expression level of at least 3 of the genes in Table 1 are determined, *e.g.*, at least 3, 5, 7 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 35, 40, 45, 50, or all 51 genes in Table 1. In particular embodiments, the expression level of at least one of ABCF2, AGMAT, BBS9, BLK, C16orf42, CAT, CDK14, CNPY4, DIP2A, ENPP4, EPX, EXOC6B, EZH1, H2AFV, HAUS2, IL1F6, IVD, MCF2L, MYL12B, MZT2B, OLFML2B, ORC6L, PDE4C, PGRMC1, RPL27A, RRAGD, SBF1, SH2B2, SNX17, SPG21, TAOK1, TEX13B, TYMP, WDR6, XRCC2, ZNF701, and ZNF721 is determined.

In certain embodiments, at least one of the genes is selected from CAT, ENPP4, RRAGD, MCF2L, SLC39A6, SNX17, PGRMC1, ZNF721, PRKAR1A, CLCN3, RNASE4, BBS9, H2AFV, POLR1B, DIP2A, EZH1, MRPL52, HAUS2, RIOK3, MZT2B, *e.g.*, the expression level of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, or all 20 genes is determined. In more particular embodiments, at least one of the genes is selected from CAT, ENPP4, RRAGD, MCF2L, SNX17, PGRMC1, ZNF721, BBS9, H2AFV, DIP2A, EZH1, HAUS2, and MZT2B, *e.g.*, the expression level of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, or all 13 genes is determined. In still more particular embodiments, the expression levels of CAT, ENPP4, RRAGD, and optionally MCF2L, are determined.

In some embodiments, an abnormal expression pattern comprises increased expression of one or more of BBS9, CAT, CLCN3, ENPP4, EZH1, H2AFV, MCF2L, PGRMC1, PRKAR1A, RNASE4, RRAGD, SLC39A6, or SNX17 (*e.g.*, 1, 2, 3, 4, 5, 67, 8, 9, 10, 11, 12, 13, or all 14), relative to a suitable control. An abnormal expression pattern may also comprise decreased expression of one or more of ABCF2, AGMAT, BLK, C16orf42, CDK14, CNPY4, DIP2A, EPX, EXOC6B, FAM128B, HAUS2, IL1F6, INHBC, IVD, MRPL52, MYL12B, OLFML2B, ORC6L, PDE4C, PGF, POLR1B, PRDX2, RIOK3, RPL27A, RPL37A, RPL38, SBF1, SH2B2, SLC35E1, SPG21, TAOK1, TEX13B, TYMP, USP34, WDR6, XRCC2, ZNF701 or ZNF721 (*e.g.*,1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, or all 37).

*Detection methods*

To determine a gene expression pattern, one or more gene expression levels are determined. Expression levels can be measured at the nucleic acid or protein level, or a combination thereof. Any means of determining gene expression levels can be employed when practicing the methods provided by the invention.

For example, nucleic acid expression levels can be determined in a number of ways including polymerase chain reaction (PCR), including reverse transcriptase (rt) PCR, real-time and quantitative PCR methods (including, *e.g.*, TAQMAN, molecular beacon, LIGHTUP, SCORPION, SIMPLEPROBES; *see, e.g.*, U.S. Pat. Nos. 5,538,848; 5,925,517; 6,174,670; 6,329,144; 6,326,145 and 6,635,427)); Northern blotting; Southern blotting of reverse transcription products and derivatives; array based methods, including blotted arrays or *in situ*-synthesized arrays; and sequencing, *e.g.*, sequencing by synthesis, pyrosequencing, dideoxy sequencing, and sequencing by ligation, or any other methods known in the art, such as discussed in Shendure *et al.*, *Nat. Rev. Genet.* 5:335-44 (2004) or Nowrousian *Euk. Cell*, 9(9): 1300–1310 (2010), including such specific platforms as HELICOS™, ROCHE™ 454, ILLUMINA™ /SOLEXA™, ABI SOLiD™, and POLONATOR™ sequencing.

Expression levels can be measured by detecting the reference nucleic acid sequences listed in Table 1, as well as complements, fragments, and similar nucleic acid sequences of the reference nucleic acid sequences listed in Table 1. "Similar nucleic acid sequences" can be naturally occurring (*e.g.*, allelic variants or homologous sequences from other species) or engineered variants to the reference nucleic acid sequences in Table 1 and will be at least about 60, 65, 70, 75, 80, 85, 90, 95, 96, 97, 98, 99% or more identical (or hybridize under highly stringent hybridization conditions to a complement of a nucleic acid sequence listed in Table 1) over a length of at least about 10, 20, 40, 60, 80, 100, 150, 200 or more nucleotides or over the entire length of the reference nucleic acid sequences in Table 1. Fragments of the reference nucleic acid sequences in Table 1—or similar nucleic acid sequences—can be of any length sufficient to distinguish the fragment from other sequences expected to be present in a mixture, *e.g.*, at least 5, 10, 15, 20, 40,

60, 80, 100, 150, 200 or more nucleotides or at least about 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 95 % of the length of the reference nucleic acid sequences in Table 1.

In particular embodiments, the expression levels of one or more of the genes in Table 1 are measured simultaneously, for example, on a nucleic acid microarray. Various microarray platforms may be adapted for use in the methods provided by the invention, including both spotted and *in situ*-synthesized arrays as well as both commercially available standard arrays as well as custom arrays, *e.g.*, custom arrays capable of detecting the expression level of one or more of the genes in Table 1. Specific microarray platforms that are useful in the methods provided by the invention are available from AFFYMETRIX™, AGILENT™ and ILLUMINA™. In more particular embodiments, the expression levels may be measured on a U133 2.0 Plus or U133 A 2.0 microarray from AFFYMETRIX™.

Protein levels can be measured by quantitative cytochemisty or histochemisty, ELISA (including direct, indirect, sandwich, competitive, multiple and portable ELISAs (see, *e.g.*, U.S. Patent No. 7,510,687)), western blotting (including one, two or higher dimensional blotting or other chromatographic means—optionally including peptide sequencing), peptide sequencing (*e.g.*, coupled to HPLC), and microarray adaptations of any of the foregoing (including antibody or protein-protein (*i.e.*, non-antibody) arrays).

Protein techniques typically, but not necessarily, employ antibodies (*e.g.*, direct sequencing). Antibodies for use in the methods provided by the invention can be directed to any of the reference peptide sequences listed in Table 1, as well as fragments of these sequences, similar peptide sequences, and fragments of similar peptide sequences. "Similar peptide sequences" can be naturally occurring (*e.g.*, allelic variants or homologous sequences from other species) or engineered variants to the reference peptide sequences in Table 1 and will exhibit substantially the same biological function and/or will be at least about 60, 65, 70, 75, 80, 85, 90, 95, 96, 97, 98, 99% or more homologous (*i.e.*, conservative substitutions (*see, e.g., Heinkoff and Heinkoff PNAS* 89 (22): 10915–10919 (1992) and Styczynski *et al., Nat. Biotech.* 26 (3): 274–275 (BLOSUM, *e.g.*, BLOSUM 45, 62 or 80) or *Dayhoff et al., Atlas of protein sequence and structure* (volume 5, supplement 3 ed.). Nat. Biomed.

Res. Found.. pp. 345–358 (PAM, *e.g.*, PAM 30 or 70)) or identical at the amino acid level over a length of at least about 10, 20, 40, 60, 80, 100, 150, 200 or more amino acids or over the entire length of the reference peptide sequences in Table 1. Fragments of the reference protein sequences in Table 1—or similar peptide sequences—can be of any length sufficient to distinguish the fragment from other sequences expected to be present in a mixture, *e.g.*, at least 5, 10, 20, 40, 60, 80, 100, 150, 200 or more amino acids or at least about 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 95 % of the length of the reference peptide sequences in Table 1.

The term "antibody," as used herein, refers to an immunoglobulin or a part thereof, and encompasses any polypeptide comprising an antigen-binding site regardless of the source, species of origin, method of production, and characteristics. As a non-limiting example, the term "antibody" includes human, orangutan, mouse, rat, goat, sheep, and chicken antibodies. The term includes but is not limited to polyclonal, monoclonal, monospecific, polyspecific, non-specific, humanized, camelized, single-chain, chimeric, synthetic, recombinant, hybrid, mutated, and CDR-grafted antibodies. For the purposes of the present invention, it also includes, unless otherwise stated, antibody fragments such as Fab, F(ab')2, Fv, scFv, Fd, dAb, VHH (also referred to as nanobodies), and other antibody fragments that retain the antigen-binding function. Antibodies also include antigen-binding molecules that are not based on immunoglobulins, as further described below.

Antibodies can be made, for example, via traditional hybridoma techniques (Kohler and Milstein, *Nature* 256: 495-499 (1975)), recombinant DNA methods (U.S. Patent No. 4,816,567), or phage display techniques using antibody libraries (Clackson *et al.*, *Nature* 352: 624-628 (1991); Marks *et al., J. Mol. Biol.* 222: 581-597 (1991)). For various other antibody production techniques, *see Antibodies: A Laboratory Manual*, eds. Harlow *et al.*, Cold Spring Harbor Laboratory, 1988.

In some embodiments, the term "antibody" includes an antigen-binding molecule based on a scaffold other than an immunoglobulin. For example, non-immunoglobulin scaffolds known in the art include small modular immunopharmaceuticals (*see, e.g.*, U.S. Patent Application Publication Nos. 20080181892 and 20080227958 published July 31, 2008 and September 18, 2008,

respectively), tetranectins, fibronectin domains (*e.g.*, AdNectins, *see* U.S. Patent Application Publication No. 2007/0082365, published April 12, 2007), protein A , lipocalins (*see, e.g.*, U.S. Patent No. 7,118,915), ankyrin repeats, and thioredoxin. Molecules based on non-immunoglobulin scaffolds are generally produced by *in vitro* selection of libraries by phage display (*see, e.g.*, Hoogenboom, *Method Mol. Biol.* 178:1-37 (2002)), ribosome display (*see, e.g.*, Hanes *et al.*, *FEBS Lett.* 450:105-110 (1999) and He and Taussig, *J. Immunol. Methods* 297:73-82 (2005)), or other techniques known in the art (*see also* Binz *et al.*, *Nat. Biotech.* 23:1257-68 (2005); Rothe *et al.*, *FASEB J.* 20:1599-1610 (2006); and U.S. Patent Nos. 7,270,950; 6,518,018; and 6,281,344) to identify high-affinity binding sequences.

Programs for sequence alignments and comparisons include FASTA (Lipman and Pearson, *Science,* 227: 1435–41 (1985) and Lipman and Pearson, *PNAS,* 85: 2444–48), BLAST (McGinnis & Madden, *Nucleic Acids Res.*, 32:W20-W25 (2004) (current BLAST reference, describing, *inter alia*, MegaBlast); Zhang *et al.*, *J. Comput. Biol.*, 7(1-2):203-14 (2000) (describing the "greedy algorithm" implemented in MegaBlast); Altschul *et al., J. Mol. Biol.*, 215:403-410 (1990) (original BLAST publication)), Needleman-Wunsch (Needleman and Wunsch, *J. Molec. Bio.*, 48 (3): 443–53(1970)), Sellers (Sellers, *Bull. Math. Biol.*, 46:501-14 (1984), and Smith-Waterman (Smith and Waterman, *J. Molec. Bio.*, 147: 195–197 (1981)), and other algorithms (including those described in Gerhard *et al., Genome Res.*, 14(10b):2121-27 (2004)), which are incorporated by reference. In particular embodiments, sequences are compared by BLAST using default parameters for nucleic acid or protein queries.

*Analysis*

Gene expression levels can be analyzed by any means in the art. Before further analysis, raw gene expression data can be transformed, *e.g.*, log-normalized, expressed as an expression ratio, *et cetera.* In particular embodiments, data may further be percentile-ranked or quantile-scaled, or modified by any nonparametric data scaling approaches.

Expression patterns can be evaluated and classified by a variety of means such as general linear model (GLM), ANOVA, regression (including logistic

regression), support vector machines (SVM), linear discriminant analysis (LDA), principal compnant analysis (PCA), k-nearest neighbor (kNN), neural network (NN), nearest mean/centroid (NM), and baysian covariate predictor (BCP). A model, such as SVM, can be developed using any of the subsets and combinations of genes described herein based on the teachings of the invention, including reference 10, below. In a particular embodiment, the SVM utilizes the expression levels of CAT, ENPP4, and RRAGD.

Suitable cutoffs for evaluating an expression pattern (*e.g.*, for classification as abnormal (cancer) or normal (non-cancer)) can be determined using routine methods such as ROC (receiver operating characteristic) analysis and adjusted to achieve the desired sensitivity (*e.g.*, at least about 50, 52, 55, 57, 60, 62, 65, 67, 70, 72, 75, 77, 80, 82, 85, 87, 90, 92, 95, 97, or 99% sensitivity) and specificity (*e.g.*, at least about 50, 52, 55, 57, 60, 62, 65, 67, 70, 72, 75, 77, 80, 82, 85, 87, 90, 92, 95, 97, or 99% specificity).

*Patients*

A patient tested and/or treated by the methods of the invention can be any human. The patient may be symptomatic or asymptomatic for lung cancer. In certain embodiments, the patient is a current or former smoker. In more particular embodiments, the patient is a former smoker. In still more particular embodiments, the patient is a former smoker who quit smoking within 10 years of isolation of the sample to be tested, for example, within about 9, 8, 7, 6, 5, 4, 3, 2.5, 2, 1.5, 1.0, 0.5 years or within about 18, 16, 14, 12, 10, 8, 6, 4, 2 or 1 months. In some embodiments, the subject is a non-smoker but may be considered to be at an increased risk for developing lung cancer—for example, the subject may have a history of exposure to second hand tobacco smoke, other environmental exposure (*e.g.*, asbestos, commercial or industrial exhaust, dust, fire and/or smoke, radiation, *et cetera*), occupational exposures, or have an increased familial risk for developing lung cancer. In certain embodiments, the patient is determined to be at risk by clinical symptoms and/or due to abnormal radiography.

A sample to be tested for abnormal gene expression can be obtained from a patient by any means known in the art from any suitable source in the oral, esophageal, nasal, and/or pulmonary system—such as lung or nasal tissues, such as epithelium. A sample can be isolated by swabbing, scraping, aspiration, lavage, sputum collection, brochoscopy, or biopsy. The sample isolation may be nondiagnostic, inconclusive, or suggestive of lung cancer—for example, based on abnormal gross tissue morphology, cytogenetic abnormalities, initial histological analysis, and/or abnormal gene expression.

*Subsequent diagnostics and treatment*

In certain embodiments, the methods provided by the invention may further comprise the step of follow-on diagnosis, for example where the subject has been identified as having an increased risk for lung cancer by methods provided by the invention. For example, in certain embodiments, the patient may be subject to techniques such as sputum cytology, flexible bronchoscopy (FB), transthoracic needle aspiration (TTNA), 18F-Fluorodeoxyglucose-positron emission tomography (FDG-PET), magnetic resonance imaging (MRI), endobronchial ultrasound and conventional or low-dose spiral computed tomography (LDCT), chest X-ray, or any biopsy. In more particular embodiments, subsequent to testing by the methods provided by the invention, the subject is determined to have lung cancer, *e.g.*, stage 0, 1a, 1b, 2a, 2b, 3a, 3b, or stage 4, as defined in the American Joint Committee on Cancer 7$^{th}$ Edition staging poster (2009), which is incorporated by reference. In still more particular embodiments, the subject is determined to have early stage lung cancer, such as stage 1a or 1b. For additional descriptions of lung cancer classifications, *see*, for example, Travis *et al.*, *J. Thorac Oncol.* 6(2); 244-285 (2011), Ravenel *et al.*, *J. Thorac Imaging* 25(4) W107-111 (2010).

Where a patient tested by the methods provided by the invention is determined to have lung cancer, suitable prophylaxis may be administered. The skilled artisan will be familiar with appropriate prophylaxis following a diagnosis of lung cancer. For example, the prophylaxis can include, *e.g.*, chemotherapy, hormonal therapy, immunotherapy (including both immunization of a patient as well

as administering antibodies to, *e.g.*, a tumor antigen or other suitable anti-neoplastic target), radiotherapy, surgery, targeted gene therapies (*e.g.*, epidermal growth factor receptor-tyrosine kinase inhibitors, such as gefitinib; and agents targeting ALK mutations and rearrangements, such as crizotinib, etc.) and combinations thereof. In particular embodiments, prophylaxis may include, in order, radiotherapy, surgery, and adjuvant therapy, such as chemotherapy.

EXEMPLIFICATION

Example 1: Initial results and materials and methods:

1.1. Gene expression data and clinical annotations

In order to identify genes with expression patterns associated with the malignant status if a patient's lung tissue, 192 Affymetrix U133a CEL files and clinical annotations were downloaded from the Boston University website: http://pulm.bumc.bu.edu/CancerDx/. The data were generated from bronchoscopy epithelial cells (BEC's) obtained from current and former smokers undergoing flexible bronchoscopy for clinical suspicion of lung cancer at four tertiary medical centers between January 2003 and April 2005. This dataset was previously published [1] and is also available in the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) with accession number GSE4115 and is referred to as the training series or validation series 1 in this report. After excluding GeneChips with fewer than 30% the total probe set detected, 150 gene expression profiles were available for analysis.

To further investigate the function of the genes selected from analysis of the BEC data, gene expression data representing lung tumor tissue and normal, non-cancerous, lung tissue was obtained from NCBI GEO with accession number GSE10072 (validation series 2). This series represented 135 fresh frozen tissue samples of adenocarcinoma and paired noninvolved lung tissue from current, former and never smokers, with biochemically validated smoking information [5]. After quality control analysis, as described in section 1.2, below, 107 gene expression profiles were available; 58 tumor and 49 non-tumor tissues, from 20 never smokers, 26 former smokers, and 28 current smokers.

Two additional validation series were gene expression profiles generated from BEC of histologically-normal individuals of varying smoking status (Validation series 3) and of individuals with dysplastic lesions, with and without lung cancer (Validation series 4). All four series used are summarized in Table 2.

Table 2: Overview of patient series used

| ID | Sample type | GEO ID | Publication Pubmed ID | Patients | Array |
|---|---|---|---|---|---|
| **Training & Validation 1** | Bronchoscopy specimens, current & former smokers, cancer, non-cancer. | GSE4115 | 17334370 | 129 | U133A |
| **Validation 2** | Lung cancer, paired normal, current, former, never smokers | GSE10072 | 18297132 | 107 | U133A |
| **Validation 3** | Bronchoscopy specimens, never, current & former smokers, non-cancer | GSE7895 | 17894889 | 104 | U133A |
| **Validation 4** | Bronchoscopy specimens, patients with dysplastic lesions | GSE12815 | 20375364 | 7 | U133A |

1.2. Data processing and quality control

Raw CEL files were obtained for each series and processed with the MAS5 algorithm (Affymetrix Inc, Santa Clara SF) before being median centered using the Affymetrix internal 100 housekeeping/reference probe set. Individual probes with intensity measurements less than 10 were excluded. Multiple probes/probe sets were reduced to one per gene symbol by using most variable probe, measured by IQR (interquartile range) across all samples and probes absent across 25% or more arrays were excluded.

Profiles from the BEC series with less than 30% of the microarray features detected were excluded from further analysis, leaving 150 which were randomly divided into a training series of 90 individuals and a validation series comprised of

the remaining 60. All 107 GeneChip profiles from the LT (lung tissue ) series (series 3 and 4) were used, as these had been previously screened for quality control issues.

All data processing was performed using Microsoft Excel (Microsoft Corporation, Redmund, WA), R (www.r-project.org), BRB ArrayTools [6] and Bioconductor [7]. Statistical analyses were performed using MedCalc 11.4 (MedCalc Software, Mariakerke, Belgium)

1.3. Gene selection & data scaling to minimize batch effects

To identify genes differentially expressed between patients with and without cancer, data from the BEC training series were analyzed with ANOVA. This method was chosen to identify individual genes expression patterns associated with cancer/no-cancer, independent to other clinical variables that may potentially impact on gene expression in BEC. The formula applied to the normalized training series dataset was:

Equation 1: ANOVA formula used to select individual genes with differential expression patterns in BEC from smokers with and without lung cancer.

$$\text{log expression} \sim [\text{cancer/no-cancer}] + [\text{pack years}] + [\text{age}] + [\text{smoking status}]$$

Where smoking status refers to a binary variable of whether an individual patient quit smoking more or less than 10 years ago and pack-years and age are also binary, depending on whether the individual is above or below the median value of these measurements.

Genes were selected based on two criteria: 1) a P-value for the difference between cancer and no cancer of $\leq 0.001$ by ANOVA and 2) an absolute fold-change of $\geq 1.5$. This ensures that each gene is both statistically and 'biologically' differentially expressed between patients with and without cancer.

Two-dimensional hierarchical clustering was performed using the gene expression data associated with the genes selected by this approach in order to visualize the differences between the two patient phenotypes described by the molecular signature. Median centering was performed on the genes, prior to

application of the average-linkage hierarchical clustering function using Cluster 3.0 [8].

Gene expression data can be affected by a number of non-biological parameters, including environmental conditions, subtle differences in wash station and scanner settings and operator error. In order to minimize the impact of systematic variation on the expression values of the selected genes, data corresponding to those passing the ANOVA selection criteria were converted to percent-rank values, which range from 0.00 to 100.00. This conversion can be performed with the function percentrank in Microsoft Excel 2007 or later (Microsoft Corporation, Redmond, WA), or with the ecdf function in the R statistical programming language.

1.4. Algorithm training & analysis of independent validation series

In order to create a diagnostic algorithm capable of predicting the malignant status of an individual smoker, a Support Vector Machine (SVM) algorithm was trained on those genes identified as being differentially expressed using ANOVA as described above [9]. A support vector machine (SVM) is a class prediction algorithm whose effectiveness has been demonstrated in many areas of machine learning research. SVMs were developed by V. Vapnik [10]

The SVM predictor is a linear function of the log-intensities or percent-rank values that best separates the data subject to penalty costs on the number of specimens misclassified. To apply the trained algorithm to a test sample, each gene expression measurement is multiplied by a specific weight, before the sum of the resulting values is taken. This value is referred to as the SVM index and is used to assign the test sample or patient to a class of tumor or no-tumor, depending upon which side of a predetermined classification threshold it lies. This threshold can be adjusted to achieve a certain sensitivity and specificity, based on the indices generated by cross validation of the training series. For the purposes of this diagnostic application, the classification threshold was tuned to give a (non-tumor result) sensitivity of $\geq$90%.

The trained SVM algorithm was then applied to the four independent validation series and the resulting SVM index and binary classification of tumor/non-tumor was recorded for each patient.

1.5. Statistical evaluation of predictions

The SVM index generated for each validation series was analyzed using Receiver Operator Curve (ROC) and Area Under the Curve (AUC) analysis. T-test, chi-square tests and General Linear Model (GLM) analysis was performed on the binary predictions of tumor/non-tumor. All T-tests were two-sided and a $P<0.05$ was considered to be statistically significant.

1.6. Results

1.6.1. Analysis of genes differentially expressed between tumor and non-tumor gene expression profiles.

After pre-processing and data normalization, ANOVA was used to compare the gene expression BEC's from patients with and without a diagnosis of lung cancer. By applying a selection criteria of $P<0.001$ and fold-change >1.5, a set of 51 unique genes was identified (Table 3). The number of genes returned by this equation by applying other selection criteria is also shown Table 3.

**Table 3: Training series ANOVA results**

| **Effect** | **Selection criteria** | **Number of genes** |
|---|---|---|
| Pack years | P<0.001 | 4 |
| Age | P<0.001 | 9 |
| Smoking status | P<0.001 | 105 |
| Cancer/no-cancer | P<0.001 | 414 |
| | **P<0.001 and fold-change > \|1.5\|** | **51** |

Hierarchical clustering was performed on the 51 genes in the 90 patient training series is shown in FIG.1. The separation of individuals with and without cancer can be seen by the structure of the horizontal dendrogram. Approximately ¾

of the 51 genes are more highly expressed (red) in individuals without cancer, compared to those with cancer.

Functional characterization of the 51 gene set was carried out using the Database for Annotation, Visualization and Integrated Discovery (DAVID ) v6.7 [11]. This method uses gene ontology, molecular pathway, chromosomal location, disease association and other data to identify significantly enriched categories of genes within a specific set. This analysis showed that the 51 gene set is enriched for genes involved in cellular response to hydrogen peroxide (enrichment score: 1.77), ribosomal structure (1.24), iron binding (0.93), hormone stimulus (0.61) and apoptosis regulation (0.59).

Ingenuity Pathway Analysis (INGENUITY SYSTEMS®; www.ingenuity.com) was also performed. As shown in FIG.2 and Table 4, the 51-gene signature exhibits significant overlap with molecular networks involved in cell-to-cell signaling, cellular growth, proliferation, tissue morphology, small molecule biochemistry, cellular development.

Table 4 Ingenuity Pathway Analysis – Significantly enriched molecular networks present in the 51-gene signature.

| **Molecules in IPA Network** | **Enrichment Score** | **Number of genes in common** | **Top Network Functions** |
|---|---|---|---|
| AGMAT, ATM, CHI3L2, CLCN3, CLEC5A, DPP7, EPX, GBP6, GBP4 (includes EG:17472), GLIPR2, IFNG, IGF1R, IKBKE, IL4, MAPK1, MYHS, MZT2B, NR3C1, OLFML2B, PDE4C, PTRH2, RAET1B, RNASE4, RPL37A (includes EG:6168), RRAGD, SBF1, SLC28A1, SREBF1, STYK1, TAOK2 (includes EG:381921), TNF, TREM3, TYMP, USP34, WDR6 (includes EG:11180) | 28 | 13 | Cell-To-Cell Signaling and Interaction, Cellular Growth and Proliferation, Gastrointestinal Disease |
| ABLIM, Akt, BLK, CAT, CDK14, CREBL2, DHRS3, DOK5, ERK1/2, ethylene glycol, EZH1, FSH, GK7P, IL1F6, Lh, LOC81691, Mapk, NFkB (complex), NRG1, NRG3, PGF, PGRMC1, PI4K2A, POP5, PRDX2, PRKAR1A, PRKX, QRFP, RIOK3, RIP, SH2B2, STK17A, TMEFF2, TP53I11, TRIB1 | 23 | 11 | Tissue Morphology, Small Molecule Biochemistry, Cellular Development |

| | | | |
|---|---|---|---|
| ABCF2, ABCF3, BBS9, BMP2K (includes EG:55589), C14ORF147, C20ORF111, CDC6, CDIPT, CWC25, DGKQ, EID1, FXYD6, HAUS2, HNF4A, INHBC, INHBE, KDM5B, MCM8, MDM2, MT1H, MT1X, ORC6L, POLR1B, Prereplicative Complex, RBL2, ROBO3, RPL27A, SLC35E1, SLC39A6, SNX17, SRP19, TMEM87A, ZNF133, ZNF701, ZNHIT6 | 21 | 11 | Genetic Disorder, Neurological Disease, Ophthalmic Disease |
| ABL1, ARHGAP26, ARHGEF4, ARHGEF25, CHCHD2, CNTNAP1, CRK/CRKL, DEF6, DOCK3, Integrin alpha 6 beta 1, LCK, LIMK2, MAPK8, MAPK15, MATK, MCF2L, MICAL1, MYL12B, NET1, PTPN18, PTPRH, PXN, RAD52, RHOA, RhoGap, RHOU, RIPK4, RPL38 (includes EG:6169), SEMA4D, SH2D2A, SPG21, TAOK1, TUB, Vegf Receptor, XRCC2 | 10 | 6 | Cellular Assembly and Organization, Cellular Movement, Cell Morphology |
| ENPP4, MECP2 | 3 | 1 | Genetic Disorder, Neurological Disease, Psychological Disorders |
| DIP2A, FSTL1 | 3 | 1 | Cardiac Infarction, Cardiovascular Disease, Genetic Disorder |
| ARL4D, CNPY4 | 3 | 1 | Molecular Transport, Protein Trafficking, Cancer |
| EXOC6B (includes EG:23233), GTP, MTERF, RALA | 2 | 1 | Gene Expression, Energy Production, Lipid Metabolism |

1.6.2. Optimization of the 51 gene SVM classification threshold

ROC analysis was performed on the cross-validated SVM indices generated for the 90 patient training series, as shown in FIG.3. An AUC of 0.891 was observed, with a 95% CI of 0.808 and 0.947 ($P<0.0001$).

To determine the optimal classification threshold value in order necessary to achieve the desired sensitivity of ≥90%, the ROC output shown in Table 5 was inspected. This represents the performance of the multi-gene classifiers at different threshold cutoff levels. The threshold of -0.16 gives a sensitivity of 90.91 (95% CI: 78.3 to 97.5) and specificity of 73.91 (95% CI: 58.9 to 85.7).

Table 5: Classification threshold variations for the 51-gene SVM and corresponding performance criteria when applied to the 90 patient training series.

| Criterion | Sensitivity | 95% CI | Specificity | 95% CI |
|---|---|---|---|---|
| < -3.038 | 0.00 | 0.0 - 8.0 | 100.00 | 92.3 - 100.0 |
| <=-1.9362 | 25.00 | 13.2 - 40.3 | 100.00 | 92.3 - 100.0 |
| <=-1.8856 | 25.00 | 13.2 - 40.3 | 97.83 | 88.5 - 99.9 |
| <=-1.5621 | 36.36 | 22.4 - 52.2 | 97.83 | 88.5 - 99.9 |
| <=-1.5584 | 36.36 | 22.4 - 52.2 | 95.65 | 85.2 - 99.5 |
| <=-1.5516 | 38.64 | 24.4 - 54.5 | 95.65 | 85.2 - 99.5 |
| <=-1.531 | 38.64 | 24.4 - 54.5 | 93.48 | 82.1 - 98.6 |
| <=-1.4931 | 50.00 | 34.6 - 65.4 | 93.48 | 82.1 - 98.6 |
| <=-1.4861 | 50.00 | 34.6 - 65.4 | 89.13 | 76.4 - 96.4 |
| <=-1.3684 | 59.09 | 43.2 - 73.7 | 89.13 | 76.4 - 96.4 |
| <=-1.2968 | 59.09 | 43.2 - 73.7 | 84.78 | 71.1 - 93.7 |
| <=-0.9718 | 77.27 | 62.2 - 88.5 | 84.78 | 71.1 - 93.7 |
| <=-0.9014 | 77.27 | 62.2 - 88.5 | 80.43 | 66.1 - 90.6 |
| <=-0.5705 | 84.09 | 69.9 - 93.4 | 80.43 | 66.1 - 90.6 |
| <=-0.563 | 84.09 | 69.9 - 93.4 | 78.26 | 63.6 - 89.1 |
| <=-0.2766 | 88.64 | 75.4 - 96.2 | 78.26 | 63.6 - 89.1 |
| <=-0.2202 | 88.64 | 75.4 - 96.2 | 73.91 | 58.9 - 85.7 |
| **<=-0.1586** | **90.91** | **78.3 - 97.5** | **73.91** | **58.9 - 85.7** |
| <=-0.0393 | 90.91 | 78.3 - 97.5 | 69.57 | 54.2 - 82.3 |
| <=-0.0027 | 93.18 | 81.3 - 98.6 | 69.57 | 54.2 - 82.3 |
| <=0.0912 | 93.18 | 81.3 - 98.6 | 67.39 | 52.0 - 80.5 |
| <=0.1059 | 95.45 | 84.5 - 99.4 | 67.39 | 52.0 - 80.5 |
| <=0.224 | 95.45 | 84.5 - 99.4 | 65.22 | 49.8 - 78.6 |
| <=0.2585 | 97.73 | 88.0 - 99.9 | 65.22 | 49.8 - 78.6 |
| <=0.3337 | 97.73 | 88.0 - 99.9 | 60.87 | 45.4 - 74.9 |
| <=0.3476 | 100.00 | 92.0 - 100.0 | 60.87 | 45.4 - 74.9 |
| <=2.0432 | 100.00 | 92.0 - 100.0 | 0.00 | 0.0 - 7.7 |

To visualize this classification threshold in the context of the 90 patient training series and associated clinical variables, a gene expression heatmap was created in which hierarchical clustering was used to order the gene expression data (columns) and patients (rows) were ordered by increasing value of the SVM index FIG.4. The subject data is also summarized in Table 5a, below.

| GWEIGHT | Cancer or normal | SVM Index | diagnostic bronch | age median | sex | Pack year Median | pack years | cancer type | cancer stage | age | race | status | Years smoking | Years since quitting |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| s215p309B1 | 1 | -3.04 | N | b | F | b | 46 | Adenocarcinoma | 3b | 86 | CAU | CC | 47 | 1 |
| s222p311B1 | 1 | -2.75 | N | b | M | b | 105 | Squamous | 2a | 70 | CAU | CC | 65 | 0 |
| s237p340B1 | 1 | -2.43 | N | b | M | b | 48 | Non Small Cell | 3a | 73 | CAU | CC | 50 | 0 |
| s258p365B1 | 1 | -2.43 | N | b | M | b | 100 | Non Small Cell | 4 | 75 | AFA | FC | 50 | 10 |
| s212p306B1 | 1 | -2.37 | Y | a | M | b | 78 | Squamous | 3a | 55 | CAU | FC | 39 | 4 |
| s213p307B1 | 1 | -2.29 | Y | a | M | b | 64 | Small Cell | Lim | 55 | CAU | CC | 44 | 0 |
| s257p364B1 | 1 | -2.21 | N | a | M | a | 25 | Large Cell | 3a | 55 | AFA | CC | 35 | 0 |
| s163p244B1 | 1 | -2.18 | N | b | M | b | 60 | Adenocarcinoma | 3b | 65 | CAU | FC | 42 | 6 |
| s193p280B1 | 1 | -2.17 | Y | a | M | a | 20 | Squamous | 3a | 52 | AFA | FC | 38 | 1 |
| s252p360B1 | 1 | -2.01 | Y | a | M | b | 75 | Adenocarcinoma | | 55 | CAU | CC | 46 | 0 |
| s206p299B1 | 1 | -1.94 | N | b | M | b | 60 | Squamous | 3a | 75 | CAU | CC | 61 | 0 |
| s220p313B1 | 0 | -1.89 | Y | a | M | a | 22 | None | None | 58 | CAU | CNC | 45 | 0 |
| s233p341B1 | 1 | -1.87 | N | b | M | b | 90 | Small Cell | Lim | 79 | CAU | CC | 61 | 0 |
| s133p186B1 | 1 | -1.72 | N | a | M | b | 50 | Adenocarcinoma | 2a | 55 | CAU | FC | 37 | 2 |
| s138p195B1 | 1 | -1.65 | N | b | M | a | 4.5 | Small Cell | Ext | 67 | CAU | FC | 6 | 42 |
| s175p260B1 | 1 | -1.60 | Y | b | F | b | 75 | Squamous | 3b | 76 | CAU | FC | 49 | 7 |
| s325p450B1 | 1 | -1.56 | Y | b | M | b | 57 | Squamous | | 83 | CAU | FC | 57 | 8 |
| s173p255B1 | 0 | -1.56 | N | b | M | a | 0.5 | None | None | 63 | ASI | FNC | 2 | 43 |
| s201p294B1 | 1 | -1.55 | N | b | M | b | 108 | ? | | 70 | CAU | FC | 55 | 7 |

| GWEIGHT | Cancer or normal | SVM Index | diagnostic bronch | age median | sex | Pack year Median | pack years | cancer type | cancer stage | age | race | status | Years smoking | Years since quitting |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| s226p319B1 | 0 | -1.53 | N | b | M | a | 25 | None | None | 69 | CAU | CNC | 51 | 0 |
| s216p310B1 | 1 | -1.50 | N | a | F | a | 32 | Adenocarcinoma | 4 | 49 | CAU | FC | 32 | 1 |
| s321p446B1 | 1 | -1.50 | N | b | F | a | 9 | Adenocarcinoma | 1a | 70 | CAU | FC | 18 | 36 |
| s142p201B1 | 1 | -1.50 | Y | b | M | b | 80 | Squamous | 4 | 61 | CAU | CC | 48 | 0 |
| s228p321B1 | 1 | -1.49 | N | b | F | a | 20 | Non Small Cell | 1b | 67 | CAU | FC | 20 | 20 |
| s328p453B1 | 1 | -1.49 | N | a | M | a | 19.5 | Adenocarcinoma | | 57 | AFA | CC | 39 | 0 |
| s195p283B1 | 0 | -1.49 | N | a | F | a | 5 | None | None | 50 | CAU | CNC | 26 | 0 |
| s301p427B1 | 0 | -1.49 | N | a | M | b | 70 | ? | | 53 | CAU | FNC | 43 | 3 |
| s271p386B1 | 1 | -1.42 | N | b | F | a | 22 | Small Cell | Lim | 72 | CAU | FC | 43 | 11 |
| s270p385B1 | 1 | -1.40 | Y | a | M | a | 29 | Squamous | 1b | 54 | CAU | CC | 45 | 0 |
| s148p211B1 | 1 | -1.38 | Y | b | F | b | 50 | Adenocarcinoma | 4 | 65 | HIS | FC | 51 | 2 |
| s154p219B1 | 1 | -1.37 | Y | a | M | a | 20 | Adenocarcinoma | 3b | 59 | CAU | FC | 20 | 26 |
| s161p231B1 | 0 | -1.34 | Y | a | M | b | 70.5 | None | None | 59 | CAU | CNC | 48 | 0 |
| s283p373B1 | 0 | -1.30 | N | a | M | a | 17 | None | None | 34 | AFA | CNC | 17 | 0 |
| s275p394B1 | 1 | -1.27 | Y | b | M | b | 71 | Small Cell | Lim | 65 | CAU | CC | 53 | 0 |
| s177p262B1 | 1 | -1.18 | Y | b | M | b | 50 | Non Small Cell | 3b | 65 | CAU | FC | 57 | 1 |
| s139p196B1 | 1 | -1.18 | Y | a | M | a | 20 | Squamous | 3a | 49 | AFA | FC | 27 | 3 |
| s209p302B1 | 1 | -1.13 | Y | b | M | a | 23 | Squamous | 4 | 80 | CAU | FC | 45 | 16 |
| s136p193B1 | 1 | -1.12 | N | b | M | b | 80 | Large Cell | 4 | 69 | AFA | CC | 51 | 0 |

| GWEIGHT | Cancer or normal | SVM Index | diagnostic bronch | age median | sex | Pack year Median | pack years | cancer type | cancer stage | age | race | status | Years smoking | Years since quitting |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| s122p171B1 | 1 | -1.09 | N | a | M | b | 120 | Small Cell | Ext | 59 | CAU | CC | 52 | 0 |
| s123p172B1 | 1 | -1.00 | N | b | M | b | 70.5 | Squamous | 1b | 64 | CAU | CC | 49 | 0 |
| s159p238B1 | 1 | -0.97 | N | b | F | b | 100 | Squamous | 3a | 69 | CAU | FC | 50 | 7 |
| s97p146B1 | 0 | -0.96 | N | a | M | a | 20.5 | None | None | 54 | AFA | CNC | 43 | 0 |
| s244p353B1 | 0 | -0.90 | N | b | M | b | 105 | ? | | 68 | CAU | CNC | 53 | 0 |
| s302p428B1 | 1 | -0.67 | Y | a | M | a | 33 | Adenocarcinoma | 4 | 46 | CAU | CC | 30 | 0 |
| s114p163B1 | 1 | -0.62 | N | b | F | b | 46 | Large Cell | 4 | 60 | CAU | FC | 46 | 2 |
| s145p205B1 | 1 | -0.57 | Y | a | F | b | 35 | Large Cell | 4 | 55 | AFA | FC | 38 | 3 |
| s151p215B1 | 0 | -0.56 | N | a | F | b | 34 | None | None | 49 | AFA | CNC | 36 | 0 |
| s255p362B1 | 1 | -0.40 | Y | a | F | b | 37 | Non Small Cell | 4 | 42 | CAU | CC | 25 | 0 |
| s274p391B1 | 1 | -0.28 | N | b | M | b | 51 | Squamous | 3b | 77 | CAU | FC | 52 | 8 |
| s172p254B1 | 0 | -0.25 | N | a | F | a | 5 | None | None | 59 | CAU | FNC | 5 | 40 |
| s300p423B1 | 0 | -0.22 | N | b | M | a | 4 | ? | | 66 | AFA | CNC | 54 | 0 |
| s144p204B1 | 1 | -0.16 | N | b | M | b | 70 | Adenocarcinoma | 1b | 68 | CAU | FC | 24 | 28 |
| s140p197B1 | 0 | -0.13 | N | b | M | b | 35 | None | None | 76 | CAU | FNC | 34 | 27 |
| s77p127B1 | 0 | -0.04 | N | a | M | a | 10 | None | None | 43 | AFA | FNC | 20 | 3 |
| s235p344B1 | 1 | 0.00 | Y | b | M | b | 53 | Squamous | 1b | 75 | CAU | FC | 52 | 6 |
| s263p377B1 | 0 | 0.09 | N | b | F | a | 24.25 | ? | | 64 | CAU | FNC | 34 | 9 |
| s150p212B1 | 1 | 0.11 | Y | b | F | b | 60 | Small Cell | Ext | 66 | CAU | FC | 30 | 17 |
| s68p118B1 | 0 | 0.22 | N | a | M | a | 28 | None | None | 38 | AFA | CNC | 30 | 0 |
| s57p113B1 | 1 | 0.26 | Y | b | M | b | 75 | Small Cell | Ext | 63 | CAU | CC | 52 | 0 |

| GWEIGHT | Cancer or normal | SVM Index | diagnostic bronch | age median | sex | Pack year Median | pack years | cancer type | cancer stage | age | race | status | Years smoking | Years since quitting |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| s292p400B1 | 0 | 0.31 | N | a | M | a | 16 | ? | | 45 | AFA | CNC | 31 | 0 |
| s113p162B1 | 0 | 0.33 | N | a | F | a | 10.5 | None | None | 36 | AFA | CNC | 23 | 0 |
| s290p404B1 | 1 | 0.35 | Y | b | M | b | 35 | Squamous | 3b | 63 | CAU | CC | 50 | 0 |
| s124p173B1 | 0 | 0.35 | N | a | M | b | 90 | None | None | 54 | CAU | CNC | 42 | 0 |
| s86p103B1 | 0 | 0.36 | N | a | M | a | 4 | None | None | 30 | HIS | CNC | 14 | 0 |
| s107p156B1 | 0 | 0.48 | N | a | M | a | 14 | None | None | 52 | AFA | FNC | | |
| s92p141B1 | 0 | 0.51 | N | a | M | a | 3.5 | None | None | 33 | ASI | CNC | 16 | 0 |
| s67p117B1 | 0 | 0.51 | N | a | M | a | 21 | None | None | 48 | AFA | CNC | 23 | 0 |
| s240p345B1 | 0 | 0.51 | N | b | M | a | 31 | None | None | 76 | CAU | FNC | 21 | 31 |
| s147p207B1 | 0 | 0.51 | N | a | M | a | 13 | None | None | 59 | HIS | FNC | 13 | 31 |
| s81p131B1 | 0 | 0.53 | N | a | M | a | 9.75 | None | None | 49 | AFA | CNC | 37 | 0 |
| s112p161B1 | 0 | 0.59 | N | b | M | a | 30 | None | None | 69 | AFA | CNC | 62 | 0 |
| s238p323B1 | 0 | 0.69 | N | b | M | b | 64 | None | None | 79 | CAU | CNC | 65 | 0 |
| s56p112B1 | 0 | 0.72 | N | b | M | b | 51 | None | None | 73 | CAU | CNC | 66 | 0 |
| s307p433B1 | 0 | 0.87 | N | a | M | b | 39.5 | ? | | 47 | AFA | CNC | 35 | 0 |
| s89p138B1 | 0 | 0.92 | N | a | F | a | 7 | None | None | 43 | AFA | FNC | 14 | 15 |
| s82p132B1 | 0 | 1.01 | N | a | F | a | 22.5 | None | None | 26 | AFA | CNC | 17 | 0 |
| s83p133B1 | 0 | 1.03 | N | a | M | a | 28.25 | None | None | 41 | CAU | CNC | 31 | 0 |
| s317p443B1 | 0 | 1.07 | Y | b | M | b | 67 | ? | | 71 | CAU | FNC | 48 | 6 |
| s109p158B1 | 0 | 1.08 | N | a | M | a | 0.5 | None | None | 49 | ASI | CNC | 7 | 0 |
| s61p108B1 | 0 | 1.20 | N | b | M | b | 56 | None | None | 71 | CAU | FNC | 56 | 3 |
| s262p376B1 | 0 | 1.24 | N | a | M | a | 5.5 | ? | | 59 | CAU | FNC | 7 | 39 |
| s106p155B1 | 0 | 1.25 | N | a | F | a | 13 | None | None | 37 | HIS | CNC | 28 | 0 |

| GWEIGHT | Cancer or normal | SVM Index | diagnostic bronch | age median | sex | Pack year Median | pack years | cancer type | cancer stage | age | race | status | Years smoking | Years since quitting |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| s110p159B1 | 0 | 1.27 | N | a | F | b | 90 | None | None | 54 | CAU | FNC | 30 | 14 |
| s309p424B1 | 0 | 1.35 | N | b | F | b | 53 | ? | | 68 | AFA | CNC | 53 | 0 |
| s88p137B1 | 0 | 1.44 | N | a | M | a | 20 | None | None | 39 | AFA | CNC | 22 | 0 |
| s250p358B1 | 0 | 1.57 | N | a | M | a | 24 | None | None | 46 | AFA | FNC | 26 | 2 |
| s75p125B1 | 0 | 1.64 | N | a | M | a | 19 | None | None | 37 | AFA | CNC | 29 | 0 |
| s98p147B1 | 0 | 1.75 | N | a | M | a | 1 | None | None | 42 | AFA | CNC | 9 | 0 |
| s119p168B1 | 0 | 1.81 | N | a | F | b | 64 | ? | | 48 | CAU | FNC | 32 | 3 |
| s84p134B1 | 0 | 2.04 | N | a | F | a | 2 | None | None | 45 | AFA | FNC | 4 | 3 |

Table 5a

1.6.3. Training series: Comparison of gene number vs. classifier performance

To determine whether a smaller number of genes could be used for classification, variations of the SVM algorithm trained on 2 to 51 genes were generated and applied to the 90-sample training series using LOOCV. The overall percent accuracy, sensitivity, specificity, positive predictive value and negative predictive values were recorded for each version of the classifier and plotted as shown in FIG.5. The mean of these five measurements was also calculated and plotted against gene number FIG.6).

This analysis revealed that with three (3) genes only, excellent LOOCV performance was achieved. Therefore acceptable performance can be achieved with minimum three gene subset of the 51 gene algorithm.

Example 2: Validation series 1: GSE4115 (BEC)

A random selection of 60 gene expression profiles of BEC's from individuals suspected of having lung cancer were partitioned from validation series 1 [1] and set aside for validation of the predictive algorithm. The 51-gene SVM was applied to the validation set, as shown in FIG.7. The results of FIG.7 are also summarized in Table 5b.

| NAME | SVM index | Cancer or normal | diagnostic bronch | age median | sex | race | Pack year Median | cancer type | cancer stage | age | smoking_status |
|---|---|---|---|---|---|---|---|---|---|---|---|
| s176p261B1 | -2.87 | 0 | N | b | M | CAU | b | None | None | 65 | quit less than 10 years |
| s188p273B1 | -2.29 | 1 | N | b | M | AFA | b | Squamous | 3b | 73 | quit for 10 years or greater |
| s197p285B1 | -2.28 | 1 | Y | a | M | CAU | a | Adenocarcinoma | 1b | 54 | quit less than 10 years |
| s266p380B1 | -2.23 | 1 | Y | a | M | CAU | b | Small Cell | Lim | 59 | quit less than 10 years |
| s219p292B1 | -2.23 | 1 | Y | b | M | CAU | b | Non Small Cell | 3b | 74 | quit for 10 years or greater |
| s199p290B1 | -2.19 | 1 | Y | a | M | CAU | b | Non Small Cell | 4 | 57 | quit less than 10 years |
| s160p239B1 | -2.11 | 1 | N | a | M | CAU | b | Squamous | 1a | 59 | quit less than 10 years |
| s241p346B1 | -2.10 | 0 | N | a | M | CAU | b | None | None | 51 | quit less than 10 years |
| s236p339B1 | -2.02 | 1 | Y | b | M | CAU | b | Small Cell | Lim | 65 | quit less than 10 years |
| s203p296B1 | -1.95 | 1 | N | b | M | CAU | b | Adenocarcinoma | 1b | 74 | quit for 10 years or greater |
| s285p398B1 | -1.89 | 1 | Y | a | M | CAU | b | Adenocarcinoma | 3b | 53 | quit less than 10 years |
| s312p435B1 | -1.82 | 1 | N | b | M | CAU | b | Small Cell | Ext | 78 | quit for 10 years or greater |
| s269p384B1 | -1.79 | 1 | Y | b | M | CAU | b | Squamous | 1b | 65 | quit less than 10 years |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| s234p343B1 | -1.78 | 1 | N | a | M | CAU | a | Squamous | 1b | 51 | quit less than 10 years |
| s223p312B1 | -1.75 | 0 | Y | b | M | CAU | b | None | None | 69 | quit less than 10 years |
| s196p284B1 | -1.74 | 0 | N | a | F | CAU | a | None | None | 53 | quit less than 10 years |
| s322p447B1 | -1.68 | 1 | N | b | F | CAU | a | Adenocarcinoma | 3b | 78 | quit for 10 years or greater |
| s239p331B1 | -1.61 | 1 | Y | a | M | CAU | a | Squamous | 3a | 58 | quit less than 10 years |
| s134p187B1 | -1.53 | 0 | N | a | M | ASI | a | None | None | 23 | quit less than 10 years |
| s286p406B1 | -1.44 | 0 | N | a | F | AFA | a | ? | | 45 | quit less than 10 years |
| s156p225B1 | -1.41 | 1 | N | b | M | CAU | b | Squamous | 4 | 68 | quit less than 10 years |
| s214p308B1 | -1.32 | 1 | Y | b | F | CAU | b | Squamous | 3a | 71 | quit less than 10 years |
| s202p295B1 | -1.24 | 1 | N | a | M | CAU | b | ? | 4 | 56 | quit less than 10 years |
| s141p198B1 | -1.22 | 0 | N | a | M | CAU | a | None | None | 55 | quit for 10 years or greater |
| s273p390B1 | -1.08 | 1 | N | b | M | CAU | a | Squamous | | 70 | quit less than 10 years |
| s190p277B1 | -0.94 | 1 | N | b | M | CAU | b | Squamous | 1b | 75 | quit less than 10 years |
| s308p434B1 | -0.92 | 1 | Y | a | M | CAU | b | Non Small Cell | 4 | 55 | quit less than 10 years |
| s162p242B1 | -0.91 | 0 | N | a | M | CAU | a | None | None | 40 | quit less than 10 years |
| s297p418B1 | -0.80 | 1 | Y | b | M | CAU | a | Small Cell | Lim | 65 | quit less than 10 years |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| s135p192B1 | -0.60 | 1 | N | b | F | CAU | b | Squamous | 4 | 68 | quit less than 10 years |
| s152p216B1 | -0.43 | 1 | Y | b | F | CAU | b | Non Small Cell | 4 | 65 | quit less than 10 years |
| s207p300B1 | -0.21 | 0 | N | b | M | CAU | b | None | None | 71 | quit less than 10 years |
| s168p249B1 | -0.14 | 1 | Y | a | M | CAU | a | Small Cell | Lim | 54 | quit less than 10 years |
| s246p355B1 | -0.10 | 0 | Y | b | M | CAU | b | None | None | 74 | quit for 10 years or greater |
| s231p328B1 | -0.07 | 1 | Y | b | M | CAU | b | Squamous | 4 | 74 | |
| s315p440B1 | -0.06 | 1 | Y | a | F | CAU | b | Squamous | 4 | 57 | quit less than 10 years |
| s304p430B1 | 0.02 | 0 | N | b | M | CAU | b | ? | | 63 | quit less than 10 years |
| s282p370B1 | 0.08 | 0 | N | b | M | CAU | b | ? | | 66 | quit less than 10 years |
| s221p314B1 | 0.14 | 1 | Y | a | M | AFA | b | Small Cell | Lim | 60 | quit less than 10 years |
| s130p183B1 | 0.18 | 1 | N | b | M | CAU | b | Adenocarcinoma | 1a | 69 | quit for 10 years or greater |
| s71p121B1 | 0.30 | 0 | N | a | M | HIS | a | None | None | 23 | quit less than 10 years |
| s116p165B1 | 0.31 | 0 | N | a | M | AFA | a | None | None | 38 | quit less than 10 years |
| s243p349B1 | 0.34 | 1 | Y | b | M | CAU | b | Non Small Cell | 4 | 83 | quit less than 10 years |
| s174p257B1 | 0.53 | 1 | N | b | M | CAU | a | Adenocarcinoma | 4 | 79 | quit for 10 years or greater |
| s299p422B1 | 0.55 | 1 | N | a | M | CAU | a | Non Small Cell | 4 | 58 | quit less than 10 years |
| s102p151B1 | 0.61 | 0 | N | a | M | CAU | a | None | None | 27 | quit less than 10 years |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| s247p356B1 | 0.64 | 0 | N | b | M | CAU | b | None | None | 74 | quit for 10 years or greater |
| s296p413B1 | 0.74 | 0 | N | b | M | CAU | b | ? | | 71 | quit less than 10 years |
| s72p122B1 | 0.79 | 0 | N | a | M | AFA | a | None | None | 38 | quit less than 10 years |
| s181p266B1 | 0.86 | 1 | Y | b | M | CAU | b | Squamous | 4 | 62 | quit for 10 years or greater |
| s58p114B1 | 0.90 | 0 | N | a | M | HIS | a | None | None | 32 | quit less than 10 years |
| s79p129B1 | 0.98 | 0 | N | a | F | AFA | a | None | None | 39 | quit less than 10 years |
| s225p318B1 | 1.01 | 0 | N | b | M | CAU | b | None | None | 71 | quit less than 10 years |
| s63p78B1 | 1.07 | 0 | N | a | F | AFA | a | None | None | 27 | quit less than 10 years |
| s73p123B1 | 1.24 | 0 | N | a | F | AFA | a | None | None | 42 | quit less than 10 years |
| s95p144B1 | 1.38 | 0 | N | a | M | AFA | a | None | None | 50 | quit for 10 years or greater |
| s104p153B1 | 1.42 | 0 | N | a | F | AFA | a | None | None | 40 | quit for 10 years or greater |
| s85p102B1 | 1.68 | 0 | N | a | M | CAU | a | None | None | 39 | quit less than 10 years |
| s120p169B1 | 1.72 | 0 | N | a | F | CAU | a | None | None | 35 | quit less than 10 years |
| s87p136B1 | 2.26 | 0 | N | a | F | AFA | a | None | None | 40 | quit less than 10 years |

Table 5b

Overall, 24/33 (73%) patients predicted to have lung cancer were correctly classified, *i.e.* received a clinical diagnosis of cancer during the length of the time in which the original study was performed. Of the 27 individuals predicted to be tumor-free, 17 were diagnosed as tumor-free (70%).

Chi-square analysis showed the association between SVM predictions and clinical diagnosis to be highly significant (P=0.001). A sensitivity of 75.00% (95% CI: 56.60% to 88.54%) and specificity of 67.86% (95% CI: 47.65% to 84.12%) were achieved.

2.1.: ROC analysis of validation series 1 and comparison to 80 gene algorithm.

ROC analysis was performed on the SVM indices generated for independent validation series 1. The association between this continuous variable and the tumor/non-tumor status of each patient was investigated with ROC analysis (FIG.8). A statistically significant AUC of 0.78 was observed, with a 95% CI of 0.65 to 0.87. This corresponds to a P-value of $< 0.0001$.

Comparison of ROC curves for the 51 gene SVM and the 81 gene 'biomarker' classifier of previously published [1] was then performed. The difference in AUC was (P=0.89).

2.2. Subset analysis of validation series 1

A number of important clinical cofactors exist within for patients included in this validation series, including smoking status (length of time between cessation of smoking and bronchoscopy procedure) and the ability of each patient to be diagnosed during the procedure itself.

To compare the performance of the 51 gene SVM in the subsets of the training series closely resembling the actual clinical need of a multi-gene diagnostic assay, performance data shown in Table 6 were calculated.

Table 6: Performance of the 51 gene SVM in subsets of validation series 1.

| **Subset of validation series 1** | **Number of tumors in predicted-tumor group (PPV)** | **Non-tumors in non-tumor group (NPV)** | **Chi-Square P-value** | **Sensitivity*** | **Specificity#** |
|---|---|---|---|---|---|
| (i) All patients | 23/31 (74%) | 20/29 | 0.001 | 71.87% | 71.43% |

| | | | | | |
|---|---|---|---|---|---|
| | | (69%) | | (53.25% to 86.25%) | (51.33% to 86.78%) |
| (ii) Quit smoking <10 years prior to bronchoscopy | 18/25 (72%) | 16/21 (76%) | 0.001 | 78.26% (56.30% to 92.54%) | 69.57% (47.08% to 86.79%) |
| (iii) Diagnosis not possible from bronchoscopy | 11/18 (61%) | 19/22 (86%) | 0.002 | 78.57% (49.20% to 95.34%) | 73.08% (52.21% to 88.43%) |
| (iv) Both criteria from (ii) and (iii) | 7/13 (54%) | 16/17 (94%) | 0.003 | 87.50% (47.35% to 99.68%) | 72.73% (49.78% to 89.27%) |

*Sensitivity: probability that a test result will be positive when the disease is present (true positive rate).

#Specificity: probability that a test result will be negative when the disease is not present (true negative rate).

The 51-gene SVM performed optimally in the subset of individuals who quit smoking less than 10-years prior to the diagnostic investigation and who could not be diagnosed by the procedure itself. In these patients the sensitivity was 88%, specificity was 73% with a negative predictive value of 94%.

The one individual incorrectly classified in the non-tumor category of subset (iv) was diagnosed with advanced (stage IV) lung cancer. A prediction of 'tumor' is less accurate in this subset, with 54% of cases in this category being diagnosed with lung cancer.

A noteworthy aspect of results from validation series 1 is the fact that 6/7 (86%) of the patients diagnosed with early-stage (1a/1b) lung cancer, were correctly assigned to the 'tumor' category. Of the 12 patients diagnosed with stage 4 cancer, six were assigned to the 'tumor' category (50%), suggesting that the 51-gene algorithm may be more accurate at correctly diagnosing patients with early-stage tumors, compared to individuals with advanced disease.

2.3. Multivariate analysis of validation series 1

General linear model analysis was performed on the 60 patients from validation series 1 to evaluate the significance of the 51 gene SVM in the context of other clinical covariates known to affect cancer risk and gene expression of BEC's from smokers. As shown in Table 7, the SVM prediction index is significantly associated with malignancy (P=0.013) in this validation series, independent to age, pack years, smoking status and sex. For comparison

purposes, the same analysis was applied to predictions made with the previously published [1] 80-gene signature. The P-value for this algorithm was also significant (P=0.029), independent to the clinical covariates; however the P-value was larger than observed with the 51 gene signature.

Table 7: General linear model comparison of association between gene expression classifiers and malignant status of lung tissue

| Covariate | P-values from General Linear Model Analysis 1: 51 gene SVM | P-values from General Linear Model Analysis 2: 80 gene algorithm |
|---|---|---|
| Age | 0.005 | 0.011 |
| Pack-years | 0.81 | 0.66 |
| Gender | 0.85 | 0.86 |
| Smoking status | 0.58 | 0.52 |
| Gene expression classifier | 0.013 | 0.029 |

In both general linear model analyses, the age of the individual was also significantly associated with malignancy, independent to the other covariates included in the model.

Example 3: Validation series 2: GSE10072 (Lung tissue)

This dataset consists of lung tumor and normal lung gene expression profiles, generated from lung tissue biopsies. This is a different tissue type than that used in the training and validation series 1, and was included in the analysis to explore the characteristics of the 51-gene signature in lung tissue directly.

3.1. Application of the 51-gene SVM to lung tumor and normal lung gene expression data.

The 51-gene SVM indices for the 25 lung tumors and 16 normal tissues (non-paired, *i.e.* from different individuals) were compared with standard Students T-Test. As shown in FIG.10, the SVM index is (on average) significantly lower in normal lung tissue, compared to lung tumors (mean indices: -1.27, -0.72 respectively; P=0.005). The dynamic range of SVM indices is also smaller in data generated from normal tissues, compared to the range of indices observed in lung tumors.

3.2. Validation Series 2: ROC Analysis of SVM index

The area under the curve was calculated for SVM indices from non-parried specimens from validation series 2 (41 gene expression profiles) as shown in FIG.11. The result of 0.76

(95% CI: 0.60 to 0.88) was statistically significant (P=0.0006), indicating the strength of the relationship between the 51-gene SVM and the malignant or non-malignant status of lung tissue.

3.3. Comparison of the 51-gene SVM index in paired lung tumor and normal tissue

Next it was investigated whether a significant difference exists in the SVM51 generated from paired lung and normal tissue gene expression profiles (*i.e.* both gene expression profiles generated from tissue material obtained from the same patient). A paired t-test was performed on the 33 pairs of tumor/normal tissue, resulting in a statistically significant difference of P=0.011. The mean index value of the normal tissues was -1.03, compared to a mean value of -0.72 from the matched tumor tissues (difference: 0.31).

Example 4: Validation series 3: GSE7895 (BEC)

This dataset consists of gene expression data generated from histologically normal BEC's of Never Smokers (n=21), Former Smokers (n=31), and Current Smokers (n=52). This series was selected in order to investigate the variation in the 51-gene SVM index between normal epithelium of individuals of varying smoking status.

A box plot and t-test of indices was performed for the three smoking-status categories (FIG.12), however no statistically significant difference was detected (P=0.25). There was a trend towards higher indices in former-smokers, compared to current smokers. As the 51-gene SVM index is positively correlated with the probability of a patient being cancer-free, the trend observed in this dataset in agreement with results observed in the training and validation series 1. It may also reflect the time-dependant, lung tissue damage repair process that occurs after an individual ceases smoking [12].

4.1. General linear model analysis of SVM index and current/former smoking status

The relationship between smoking status of the histologically-normal individuals and the value of the 51-gene SVM index was further explored using multivariate, general linear model analysis. As shown in FIG.13, when analyzed in a setting where variation in age and pack-years are taken into account, a statistically significant difference between current and former smokers is observed (P=0.042).

The mean (adjusted) difference between the indices of former and current smokers is 0.54 (standard error: 0.27), with the former smokers in this validation series exhibiting a small but significantly larger (*i.e.* more 'tumor-like') SVM index compared to current smokers.

Example 5: Validation Series 4: GSE12815 (BEC)

This validation set consisted of gene expression profiles generated from BEC's from two smokers with COPD (no lung cancer), one smoker without COPD (no lung cancer), two patients with lung cancer and two individuals without lung cancer. As this validation series is too small for meaningful statistical analysis, the results of the 51-gene SVM analysis are shown in Table 8.

Table 8: Patient summary and 51-gene SVM predictions for validation series 5

| GeneChip ID | Sample ID | Smoking status | Years since quitting | SVM Index | SVM Prediction | Cancer status |
|---|---|---|---|---|---|---|
| GSM321999 | BT5 | Current Smoker; 20 pyers; 60-64 years old | 0 | -0.88 | Cancer | Cancer |
| GSM322000 | BT16 | Former Smoker; 21 pyers; 50-54 years old | 0.08 | -1.17 | Cancer | No cancer |
| GSM322003 | XXX | Current smoker, COPD | 0 | -1.78 | Cancer | No cancer |
| GSM322004 | 80 | Current smoker, COPD | 0 | 2.20 | **Non-cancer** | **No cancer** |
| GSM322005 | 327 | Current smoker , no COPD | 0 | 2.30 | **Non-cancer** | **No cancer** |
| GSM322001 | BT17 | Former Smoker; 58 pyers; 70-74 years old | **30** | -1.35 | Cancer | Cancer |
| GSM322002 | BT18 | Former Smoker; 15 pyers; 45-49 years old | **20** | -1.84 | Cancer | No cancer |

Two current-smoking patients (ID's: 80 and 327) were classified as 'non-tumor' by the 51-gene SVM, which agreed with the clinical diagnosis. Three other currently-smoking patients (BT5, BT16, XXX) were classified as 'tumor', however only BT5 was diagnosed with lung cancer.

Two other patients (BT17 and BT18) quit smoking 30 and 20 years prior to their bronchoscopy procedures, respectively. Patient BT17 was predicted to have lung cancer, in

agreement with their clinical diagnosis. Patient BT18 was also predicted to have lung cancer, however had not been diagnosed with this condition.

The results of this validation series are agreement with the previous analyses which show the 51-gene SVM to be more accurate in diagnosing lung tumors in patients who quit smoking less than 10 years prior to the diagnostic investigation. Furthermore, both patients predicted to be cancer-free, were actually free of lung cancer, in keeping with the high negative predictive value of the algorithm observed in validation series 1.

A multi-gene diagnostic assay has been developed using AFFYMETRIX™ whole-genome profiles generated from BEC of individuals who were suspected of having lung cancer. By selecting genes with robust patterns of differential expression between individuals who were confirmed (with traditional diagnostic methods) to have lung cancer and those who were not, a 51-gene signature was created. These genes were used to train a SVM algorithm, capable of classifying a new sample as either tumor or non-tumor, optimized for high sensitivity and negative predictive value.

The classifier was applied to four separate independent validation series and was shown to be highly accurate at identifying individuals without lung cancer, independent to other clinical variables known to influence a smoker's risk of lung cancer, such as age, pack-years, current smoking status and gender. The assay is also accurate at identifying patients with early stage cancer, who quit smoking within 10 years prior to collecting the BEC's for diagnostic analysis. In this subset of individuals, the 51-gene signature is able to identify patients who have a clinical suspicion of lung cancer who are free of lung cancer with 94% accuracy (NPV).

The 51-gene SVM index was also shown to be significantly associated with malignancy when calculated using gene expression data generated from biopsies of lung cancer and normal lung tissue, indicating the utility of the assay in a different tissue-type than it was developed from.

In a multivariate analysis of histologically-normal BEC gene expression data, the 51-gene SVM index was significantly different between current and former smokers, potentially reflecting a reversal of tissue damage that occurs when an individual ceases smoking. Table 9 Summarizes the SVM results

Table 9: Genes present in the 51 gene SVM algorithm. Probe ID and gene annotation obtained from BioConductor Affymetrix annotation packages.

| Affymetrix ProbeSet ID | Gene name | Accession | Unigene ID | Gene Symbol | Cytoband | SWM Weight |
|---|---|---|---|---|---|---|
| 200605_s_at | protein kinase, cAMP-dependent, regulatory, type I, alpha (tissue specific extinguisher 1) | NM_002734 | Hs.280342 | PRKAR1A | 17q23-q24 | -0.0991 |
| 200991_s_at | sorting nexin 17 | NM_014748 | Hs.278569 | SNX17 | 2p23-p22 | -0.4909 |
| 201120_s_at | progesterone receptor membrane component 1 | AL547946 | Hs.90061 | PGRMC1 | Xq22-q24 | -0.0763 |
| 201432_at | catalase | NM_001752 | Hs.502302 | CAT | 11p13 | -0.6615 |
| 201732_s_at | chloride channel 3 | AF029346 | Hs.481186 | CLCN3 | 4q33 | -1.3756 |
| 202089_s_at | solute carrier family 39 (zinc transporter), member 6 | NM_012319 | Hs.725276 | SLC39A6 | 18q12.2 | -0.1778 |
| 203249_at | enhancer of zeste homolog 1 (Drosophila) | AB002386 | Hs.194669 | EZH1 | 17q21.1-q21.3 | -1.3011 |
| 204160_s_at | ectonucleotide pyrophosphatase/phosphodiesterase 4 (putative) | AW194947 | Hs.643497 | ENPP4 | 6p21.1 | -0.3385 |
| 205158_at | ribonuclease, RNase A family, 4 | NM_002937 | Hs.283749 | RNASE4 | 14q11.1 | 0.1278 |
| 205367_at | SH2B adaptor protein 2 | NM_020979 | Hs.489448 | SH2B2 | 7q22 | 0.0529 |
| 206792_x_at | phosphodiesterase 4C, cAMP-specific | NM_000923 | Hs.132584 | PDE4C | 19p13.11 | -0.1521 |
| 207365_x_at | ubiquitin specific peptidase 34 | NM_014709 | Hs.644708 | USP34 | 2p15 | -0.0769 |
| 207598_x_at | X-ray repair complementing defective repair in Chinese hamster cells 2 | NM_005431 | Hs.647093 | XRCC2 | 7q36.1 | -0.1071 |
| 207688_s_at | inhibin, beta C | NM_005538 | Hs.632722 | INHBC | 12q13.1 | 0.6389 |
| 209246_at | ATP-binding cassette, sub-family F (GCN20), member 2 | AF261091 | Hs.654958 | ABCF2 | 7q36 | 0.7865 |
| 209958_s_at | Bardet-Biedl syndrome 9 | AF095771 | Hs.372360 | BBS9 | 7p14 | -0.4511 |
| 210934_at | B lymphoid tyrosine kinase | BC004473 | Hs.146591 | BLK | 8p23-p22 | -0.0373 |
| 211502_s_at | cyclin-dependent kinase 14 | AF119833 | Hs.430742 | CDK14 | 7q21-q22 | 0.8952 |
| 212044_s_at | ribosomal protein L27a | BE737027 | Hs.523463 | RPL27A | 11p15 | -0.2036 |
| 212206_s_at | H2A histone family, member V | BF343852 | Hs.488189 | H2AFV | 7p13 | 0.348 |
| 212935_at | MCF.2 cell line derived transforming sequence-like | AB002360 | Hs.170422 | MCF2L | 13q34 | -1.4179 |
| 213105_s_at | chromosome 16 open reading frame 42 | AI799802 | Hs.134846 | C16orf42 | 16p13.3 | 0.5584 |
| 213125_at | olfactomedin-like 2B | AW007573 | Hs.507515 | OLFML2B | 1q23.3 | -0.3557 |
| 213383_at | SET binding factor 1 | AW593269 | Hs.589924 | SBF1 | 22q13.33 | 0.7841 |
| 214041_x_at | ribosomal protein L37a | BE857772 | Hs.433701 | RPL37A | 2q35 | -0.5478 |
| 214627_at | eosinophil peroxidase | X14346 | Hs.279259 | EPX | 17q23.1 | -0.1525 |
| 215067_x_at | peroxiredoxin 2 | AU147942 | Hs.432121 | PRDX2 | 19p13.2 | 0.2413 |
| 215179_x_at | placental growth factor | AK023843 | Hs.252820 | PGF | 14q24.3 | -0.0949 |
| 215383_x_at | spastic paraplegia 21 (autosomal recessive, Mast syndrome) | AL137312 | Hs.242458 | SPG21 | 15q21-q22 | 0.4007 |

| Affymetrix ProbeSet ID | Gene name | Accession | Unigene ID | Gene Symbol | Cytoband | SWM Weight |
|---|---|---|---|---|---|---|
| 215417_at | exocyst complex component 6B | AI803703 | Hs.303454 | EXOC6B | 2p13.2 | 0.8071 |
| 215529_x_at | DIP2 disco-interacting protein 2 homolog A (Drosophila) | AI590053 | Hs.189585 | DIP2A | 21q22.3 | 0.152 |
| 215588_x_at | RIO kinase 3 (yeast) | AK024958 | Hs.445511 | RIOK3 | 18q11.2 | 0.6296 |
| 215978_x_at | zinc finger protein 721 | AK021514 | Hs.428360 | ZNF721 | 4p16.3 | 0.2899 |
| 216310_at | TAO kinase 1 | AK024376 | Hs.631758 | TAOK1 | 17q11.2 | 0.0602 |
| 216495_x_at | isovaleryl-CoA dehydrogenase | AF043584 | Hs.513646 | IVD | 15q14-q15 | 0.8668 |
| 217497_at | thymidine phosphorylase | AW613387 | Hs.592212 | TYMP | 22q13 | 1.0144 |
| 217700_at | canopy 4 homolog (zebrafish) | AI818951 | Hs.632293 | CNPY4 | 7q22.1 | 0.0726 |
| 217734_s_at | WD repeat domain 6 | NM_018031 | Hs.654815 | WDR6 | 3p21.31 | 0.4187 |
| 219105_x_at | origin recognition complex, subunit 6 like (yeast) | NM_014321 | Hs.49760 | ORC6L | 16q12 | -0.1093 |
| 219792_at | agmatine ureohydrolase (agmatinase) | NM_024758 | Hs.567583 | AGMAT | 1p36.21 | 0.4859 |
| 220071_x_at | HAUS augmin-like complex, subunit 2 | NM_018097 | Hs.14347 | HAUS2 | 15q15.2 | -0.0301 |
| 220113_x_at | polymerase (RNA) I polypeptide B, 128kDa | NM_019014 | Hs.86337 | POLR1B | 2q13 | 0.1501 |
| 220242_x_at | zinc finger protein 701 | NM_018260 | Hs.235167 | ZNF701 | 19q13.41 | 0.1925 |
| 220720_x_at | mitotic spindle organizing protein 2B | NM_025029 | Hs.469925 | MZT2B | 2q21.1 | -0.3857 |
| 220796_x_at | solute carrier family 35, member E1 | NM_024881 | Hs.620596 | SLC35E1 | 19p13.11 | 0.0713 |
| 221034_s_at | testis expressed 13B | NM_031273 | Hs.333130 | TEX13B | Xq22.3 | 0.5532 |
| 221404_at | interleukin 1 family, member 6 (epsilon) | NM_014440 | Hs.278910 | IL1F6 | 2q12-q14.1 | 0.1517 |
| 221474_at | myosin, light chain 12B, regulatory | U26162 | Hs.190086 | MYL12B | 18p11.31 | 0.2217 |
| 221524_s_at | Ras-related GTP binding D | AF272036 | Hs.31712 | RRAGD | 6q15-q16 | -0.7208 |
| 221943_x_at | ribosomal protein L38 | AW303136 | Hs.380953 | RPL38 | 17q25.1 | -0.3581 |
| 221997_s_at | mitochondrial ribosomal protein L52 | AI560951 | Hs.355935 | MRPL52 | 14q11.2 | 0.7176 |

Table 10 summarizes additional data (log normalized and percentile-ranked) about the genes used to generate the SVM.

Table 10

| ProbeSet | Symbol | Parametric p-value | t-value | Non-Tumor (Mean) | Tumor (Mean) | Fold-change |
|---|---|---|---|---|---|---|
| 204160_s_at | ENPP4 | 0.0000001 | -5.722 | 1.6 | 1.77 | 0.9 |
| 201432_at | CAT | 0.0000001 | -6.195 | 1.82 | 1.95 | 0.94 |
| 221524_s_at | RRAGD | 0.0000002 | -5.634 | 1.42 | 1.59 | 0.89 |
| 212935_at | MCF2L | 0.0000013 | -5.195 | 1.1 | 1.27 | 0.87 |
| 202089_s_at | SLC39A6 | 0.0000046 | -4.884 | 1.27 | 1.44 | 0.88 |

| ProbeSet | Symbol | Parametric p-value | t-value | Non-Tumor (Mean) | Tumor (Mean) | Fold-change |
|---|---|---|---|---|---|---|
| 200991_s_at | SNX17 | 0.0000064 | -4.8 | 1.34 | 1.54 | 0.87 |
| 201120_s_at | PGRMC1 | 0.0000098 | -4.698 | 1.3 | 1.52 | 0.85 |
| 215978_x_at | ZNF721 | 0.0000106 | 4.674 | 1.78 | 1.69 | 1.05 |
| 200605_s_at | PRKAR1A | 0.0000111 | -4.662 | 1.73 | 1.88 | 0.92 |
| 201732_s_at | CLCN3 | 0.000012 | -4.681 | 1.11 | 1.25 | 0.89 |
| 205158_at | RNASE4 | 0.0000211 | -4.495 | 1.47 | 1.65 | 0.89 |
| 209958_s_at | BBS9 | 0.0000329 | -4.386 | 1.11 | 1.26 | 0.88 |
| 212206_s_at | H2AFV | 0.0000396 | -4.328 | 1.12 | 1.26 | 0.89 |
| 220113_x_at | POLR1B | 0.0000511 | 4.26 | 1.55 | 1.49 | 1.04 |
| 215529_x_at | DIP2A | 0.0000664 | 4.189 | 1.38 | 1.28 | 1.08 |
| 203249_at | EZH1 | 0.0000988 | -4.108 | 1.21 | 1.35 | 0.9 |
| 221997_s_at | MRPL52 | 0.0001043 | 4.065 | 1.31 | 1.23 | 1.06 |
| 215588_x_at | RIOK3 | 0.0001322 | 3.999 | 1.79 | 1.72 | 1.04 |
| 220071_x_at | HAUS2 | 0.0001322 | 3.999 | 1.78 | 1.72 | 1.04 |
| 220720_x_at | MZT2B | 0.0001648 | 3.937 | 1.6 | 1.49 | 1.07 |
| 205367_at | SH2B2 | 0.000218 | 3.857 | 1.6 | 1.54 | 1.04 |
| 220242_x_at | ZNF701 | 0.000279 | 3.788 | 1.17 | 1.11 | 1.05 |
| 213105_s_at | C16orf42 | 0.000366 | 3.716 | 1.36 | 1.27 | 1.08 |
| 207688_s_at | INHBC | 0.0003682 | 3.705 | 1.24 | 1.17 | 1.06 |
| 215067_x_at | PRDX2 | 0.0004099 | 3.674 | 1.55 | 1.46 | 1.06 |
| 207598_x_at | XRCC2 | 0.0004774 | 3.629 | 1.39 | 1.32 | 1.06 |
| 216495_x_at | IVD | 0.000605 | 3.558 | 1.45 | 1.35 | 1.07 |
| 216310_at | TAOK1 | 0.0006626 | 3.53 | 1.53 | 1.45 | 1.06 |
| 220796_x_at | SLC35E1 | 0.0009844 | 3.409 | 1.86 | 1.8 | 1.03 |
| 215383_x_at | SPG21 | 0.0015874 | 3.26 | 1.57 | 1.5 | 1.05 |
| 206792_x_at | PDE4C | 0.001972 | 3.19 | 1.96 | 1.93 | 1.02 |
| 219792_at | AGMAT | 0.0038212 | 2.972 | 1.21 | 1.16 | 1.04 |
| 221404_at | IL1F6 | 0.0043026 | 2.935 | 1.17 | 1.11 | 1.05 |
| 215417_at | EXOC6B | 0.0045969 | 2.933 | 1.09 | 1.05 | 1.04 |
| 213383_at | SBF1 | 0.0047047 | 2.901 | 1.25 | 1.19 | 1.05 |
| 217497_at | TYMP | 0.0061321 | 2.826 | 1.09 | 1.05 | 1.04 |
| 210934_at | BLK | 0.0063738 | 2.796 | 1.2 | 1.14 | 1.05 |
| 221034_s_at | TEX13B | 0.0077024 | 2.73 | 1.2 | 1.14 | 1.05 |
| 211502_s_at | CDK14 | 0.0081815 | 2.726 | 1.09 | 1.05 | 1.04 |
| 215179_x_at | PGF | 0.0089887 | 2.672 | 1.77 | 1.74 | 1.02 |
| 219105_x_at | ORC6L | 0.0145461 | 2.494 | 1.2 | 1.15 | 1.04 |
| 217734_s_at | WDR6 | 0.0165359 | 2.444 | 1.55 | 1.49 | 1.04 |
| 217700_at | CNPY4 | 0.0191607 | 2.388 | 1.17 | 1.12 | 1.05 |
| 207365_x_at | USP34 | 0.0264526 | 2.257 | 1.79 | 1.75 | 1.02 |
| 214627_at | EPX | 0.032782 | 2.172 | 1.23 | 1.17 | 1.04 |

| ProbeSet | Symbol | Parametric p-value | t-value | Non-Tumor (Mean) | Tumor (Mean) | Fold-change |
|---|---|---|---|---|---|---|
| 209246_at | ABCF2 | 0.0362803 | 2.13 | 1.16 | 1.11 | 1.05 |
| 214041_x_at | RPL37A | 0.0450646 | 2.033 | 1.81 | 1.74 | 1.04 |
| 213125_at | OLFML2B | 0.0681089 | 1.848 | 1.12 | 1.09 | 1.03 |
| 212044_s_at | RPL27A | 0.0762717 | 1.794 | 1.58 | 1.51 | 1.04 |
| 221474_at | MYL12B | 0.2770767 | 1.094 | 1.98 | 1.97 | 1 |
| 221943_x_at | RPL38 | 0.5045344 | 0.67 | 1.88 | 1.87 | 1.01 |

References:

1. Spira, A., et al., Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. Nat Med, 2007. **13**(3): p. 361-6.

2. Gordon, G.J., et al., Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Res, 2002. **62**(17): p. 4963-4967.

3. Cerutti, J.M., et al., A preoperative diagnostic test that distinguishes benign from malignant thyroid carcinoma based on gene expression. J Clin Invest, 2004. **113**(8): p. 1234-1242.

4. Bridgewater, J., et al., Gene expression profiling may improve diagnosis in patients with carcinoma of unknown primary. Br J Cancer, 2008. **98**(8): p. 1425-30.

5. Landi, M.T., et al., Gene Expression Signature of Cigarette Smoking and Its Role in Lung Adenocarcinoma Development and Survival. PLoS ONE, 2008. **3**(2): p. e1651.

6. Simon, R., et al., *Analysis of Gene Expression Data Using BRB-Array Tools.* Cancer Inform, 2007. **3**: p. 11-7.

7. Gentleman, R.C., et al., Bioconductor: open software development for computational biology and bioinformatics. Genome Biol, 2004. **5**(10): p. R80.

8. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14863-14868.

9. Yeang, C.H., et al., *Molecular classification of multiple tumor types.* Bioinformatics, 2001. **17 Suppl 1**: p. S316-S322.

10. Vapnik, V., *Statistical Learning Theory.* 1998, New York: John Wiley.

11. Huang, D.W., B.T. Sherman, and R.A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protocols, 2008. **4**(1): p. 44-57.

12. Beane, J., et al., Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. Genome Biol, 2007. **8**(9): p. R201.

It should be understood that for all numerical bounds describing some parameter in this application, such as "about," "at least," "less than," and "more than," the description also necessarily encompasses any range bounded by the recited values. Accordingly, for example, the description at least 1, 2, 3, 4, or 5 also describes, *inter alia*, the ranges 1-2, 1-3, 1-4, 1-5, 2-3, 2-4, 2-5, 3-4, 3-5, and 4-5, *et cetera*.

For all patents, applications, or other reference cited herein, such as non-patent literature and reference sequence information, it should be understood that it is incorporated by reference in its entirety for all purposes as well as for the proposition that is recited. Where any conflict exits between a document incorporated by reference and the present application, this application will control. All information associated with reference gene sequences disclosed in this application, such as GeneIDs or accession numbers, including, for example, genomic loci, genomic sequences, functional annotations, allelic variants, and reference mRNA (including, *e.g.*, exon boundaries or response elements) and protein sequences (such as conserved domain structures) are hereby incorporated by reference in their entirety.

While this invention has been particularly shown and described with references to example embodiments thereof, it will be understood by those skilled in the art that various changes in form and details may be made therein without departing from the scope of the invention encompassed by the appended claims.

CLAIMS

What is claimed is:

1. A method of predicting the risk of lung cancer in a patient comprising:

a) testing an isolated sample from a patient who is a former cigarette smoker and ceased smoking within 10 years of isolation of the sample, wherein the sample is tested for an abnormal expression pattern of three or more genes selected from the group consisting of: ABCF2, AGMAT, BBS9, BLK, C16orf42, CAT, CDK14, CLCN3, CNPY4, DIP2A, ENPP4, EPX, EXOC6B, EZH1, H2AFV, HAUS2, IL1F6, INHBC, IVD, MCF2L, MRPL52, MYL12B, MZT2B, OLFML2B, ORC6L, PDE4C, PGF, PGRMC1, POLR1B, PRDX2, PRKAR1A, RIOK3, RNASE4, RPL27A, RPL37A, RPL38, RRAGD, SBF1, SH2B2, SLC35E1, SLC39A6, SNX17, SPG21, TAOK1, TEX13B, TYMP, USP34, WDR6, XRCC2, ZNF701, and ZNF721,

wherein an abnormal expression pattern of the three or more genes predicts an increased risk of lung cancer in the patient.

2. The method of Claim 1, wherein at least one of the three or more genes is selected from the group consisting of ABCF2, AGMAT, BBS9, BLK, C16orf42, CAT, CDK14, CNPY4, DIP2A, ENPP4, EPX, EXOC6B, EZH1, H2AFV, HAUS2, IL1F6, IVD, MCF2L, MYL12B, MZT2B, OLFML2B, ORC6L, PDE4C, PGRMC1, RPL27A, RRAGD, SBF1, SH2B2, SNX17, SPG21, TAOK1, TEX13B, TYMP, WDR6, XRCC2, ZNF701, and ZNF721.

3. A method of predicting the risk of lung cancer in a patient comprising:

a) testing an isolated sample from the patient, wherein the sample is tested for an abnormal expression pattern of three or more genes selected from the group consisting of: ABCF2, AGMAT, BBS9, BLK, C16orf42, CAT, CDK14, CLCN3, CNPY4, DIP2A, ENPP4, EPX, EXOC6B, EZH1, H2AFV, HAUS2, IL1F6, INHBC, IVD, MCF2L, MRPL52, MYL12B, MZT2B, OLFML2B, ORC6L, PDE4C, PGF, PGRMC1, POLR1B, PRDX2, PRKAR1A, RIOK3, RNASE4, RPL27A, RPL37A, RPL38, RRAGD, SBF1, SH2B2,

SLC35E1, SLC39A6, SNX17, SPG21, TAOK1, TEX13B, TYMP, USP34, WDR6, XRCC2, ZNF701, and ZNF721,

wherein an abnormal expression pattern of the three or more genes predicts an increased risk of lung cancer in the patient, wherein at least one of the at least three genes is selected from the group consisting of ABCF2, AGMAT, BBS9, BLK, C16orf42, CAT, CDK14, CNPY4, DIP2A, ENPP4, EPX, EXOC6B, EZH1, H2AFV, HAUS2, IL1F6, IVD, MCF2L, MYL12B, MZT2B, OLFML2B, ORC6L, PDE4C, PGRMC1, RPL27A, RRAGD, SBF1, SH2B2, SNX17, SPG21, TAOK1, TEX13B, TYMP, WDR6, XRCC2, ZNF701, and ZNF721.

4. The method of Claim 3, wherein the patient is a former cigarette smoker and ceased smoking within 10 years of isolation of the sample.

5. The method of any one of Claims 1-4, wherein the patient had a nondiagnostic bronchoscopy for lung cancer.

6. The method of any one of Claims 1-5, wherein the abnormal expression pattern of the three or more genes comprises increased expression of one or more of BBS9, CAT, CLCN3, ENPP4, EZH1, H2AFV, MCF2L, PGRMC1, PRKAR1A, RNASE4, RRAGD, SLC39A6, or SNX17.

7. The method of any one of Claims 1-6, wherein the abnormal expression pattern of the three or more genes comprises decreased expression of one or more of ABCF2, AGMAT, BLK, C16orf42, CDK14, CNPY4, DIP2A, EPX, EXOC6B, FAM128B, HAUS2, IL1F6, INHBC, IVD, MRPL52, MYL12B, OLFML2B, ORC6L, PDE4C, PGF, POLR1B, PRDX2, RIOK3, RPL27A, RPL37A, RPL38, SBF1, SH2B2, SLC35E1, SPG21, TAOK1, TEX13B, TYMP, USP34, WDR6, XRCC2, ZNF701 or ZNF721.

8. The method of any one of Claims 1-5, wherein the abnormal expression pattern of the three or more genes comprises:

a) increased expression of one or more of BBS9, CAT, CLCN3, ENPP4, EZH1, H2AFV, MCF2L, PGRMC1, PRKAR1A, RNASE4, RRAGD, SLC39A6, or SNX17; and

b) decreased expression of one or more of ABCF2, AGMAT, BLK, C16orf42, CDK14, CNPY4, DIP2A, EPX, EXOC6B, FAM128B, HAUS2, IL1F6, INHBC, IVD, MRPL52, MYL12B, OLFML2B, ORC6L, PDE4C, PGF, POLR1B, PRDX2, RIOK3, RPL27A, RPL37A, RPL38, SBF1, SH2B2, SLC35E1, SPG21, TAOK1, TEX13B, TYMP, USP34, WDR6, XRCC2, ZNF701 or ZNF721.

9. The method of any one of Claims 1-8, wherein at least one of the three or more genes is selected from CAT, ENPP4, RRAGD, MCF2L, SLC39A6, SNX17, PGRMC1, ZNF721, PRKAR1A, CLCN3, RNASE4, BBS9, H2AFV, POLR1B, DIP2A, EZH1, MRPL52, HAUS2, RIOK3, MZT2B.

10. The method of Claim 9, wherein at least one of the three or more genes is selected from CAT, ENPP4, RRAGD, MCF2L, SNX17, PGRMC1, ZNF721, BBS9, H2AFV, DIP2A, EZH1, HAUS2, and MZT2B.

11. The method of any one of Claims 1-10, wherein the abnormal expression pattern of the three or more genes is measured at the protein level.

12. The method of any one of Claims 1-10, wherein the abnormal expression pattern of the three or more genes is measured at the nucleic acid level.

13. The method of any one of Claims 1-12, wherein the abnormal expression pattern of the three or more genes is measured simultaneously.

14. The method of any one of Claims 1-13, wherein the abnormal expression pattern of the three or more genes is measured simultaneously on a microarray.

15. The method of any one of Claims 1-14, wherein the abnormal expression of the three or more genes is measured at the nucleic acid level on a microarray and the microarray is a Exon 1.0 ST, Gene 1.0 ST, U 95, U133, U133A 2.0, or U133 Plus 2.0.

16. The method of any one of Claims 1-15, wherein the expression level of at least 5, 7, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 35, 40, 45, 50, or all 51 genes is determined.

17. The method of any one of Claims 1-16, wherein the expression levels are log-normalized.

18. The method of Claim 17, wherein the log-normalized expression levels are converted to percent-rank values.

19. The method of any one of Claims 1-18, further comprising the step of determining a disease index for the individual using the expression pattern of three or more genes.

20. The method of Claim 19, wherein the disease index is a support vector machine (SVM) index.

21. The method of Claim 20, wherein the SVM index is determined using SVM weights substantially similar to MCF2L (-1.4179), CLCN3 (-1.3756), EZH1 (-1.3011), TYMP (1.0144), CDK14 (0.8952), IVD (0.8668), EXOC6B (0.8071), ABCF2 (0.7865), SBF1 (0.7841), RRAGD (-0.7208), MRPL52 (0.7176), CAT (-0.6615), INHBC (0.6389), RIOK3 (0.6296), C16orf42 (0.5584), TEX13B (0.5532), RPL37A (-0.5478), SNX17 (-0.4909), AGMAT (0.4859), BBS9 (-0.4511), WDR6 (0.4187), SPG21 (0.4007), MZT2B (-0.3857), RPL38 (-0.3581), OLFML2B (-0.3557), H2AFV (0.348), ENPP4 (-0.3385), ZNF721 (0.2899), PRDX2 (0.2413), MYL12B (0.2217), RPL27A (-0.2036), ZNF701 (0.1925), SLC39A6 (-0.1778), EPX (-0.1525), PDE4C (-0.1521), DIP2A (0.152), IL1F6 (0.1517), POLR1B (0.1501), RNASE4 (0.1278), ORC6L (-0.1093), XRCC2 (-0.1071), PRKAR1A (-0.0991), PGF (-0.0949), USP34 (-0.0769), PGRMC1 (-0.0763), CNPY4 (0.0726), SLC35E1 (0.0713), TAOK1 (0.0602), SH2B2 (0.0529), BLK (-0.0373), and HAUS2 (-0.0301).

22. The method of any one of Claims 1-21, wherein the isolated sample comprises lung epithelial tissue.

23. The method of Claim 22, wherein the sample is obtained by bronchoscopy.

24. The method of any one of Claims 1-21, wherein the isolated sample comprises nasal epithelial tissue.

25. The method of any one of Claims 1-24, further comprising the step of one or more of non-invasive imaging, lung mass biopsy, or histological and/or cytological analysis of an isolated sample from the oral, esophageal, nasal, and/or pulmonary system of the subject.

26. A kit comprising reagents for detecting the expression level of three or more genes selected from the group consisting of: ABCF2, AGMAT, BBS9, BLK, C16orf42, CAT, CDK14, CLCN3, CNPY4, DIP2A, ENPP4, EPX, EXOC6B, EZH1, H2AFV, HAUS2, IL1F6, INHBC, IVD, MCF2L, MRPL52, MYL12B, MZT2B, OLFML2B, ORC6L, PDE4C, PGF, PGRMC1, POLR1B, PRDX2, PRKAR1A, RIOK3, RNASE4, RPL27A, RPL37A, RPL38, RRAGD, SBF1, SH2B2, SLC35E1, SLC39A6, SNX17, SPG21, TAOK1, TEX13B, TYMP, USP34, WDR6, XRCC2, ZNF701, and ZNF721, for detecting lung cancer in a patient; the kit optionally containing suitable positive controls.

27. A set of oligonucleotide primers for detecting the nucleic acid expression level three or more genes selected from the group consisting of: ABCF2, AGMAT, BBS9, BLK, C16orf42, CAT, CDK14, CLCN3, CNPY4, DIP2A, ENPP4, EPX, EXOC6B, EZH1, H2AFV, HAUS2, IL1F6, INHBC, IVD, MCF2L, MRPL52, MYL12B, MZT2B, OLFML2B, ORC6L, PDE4C, PGF, PGRMC1, POLR1B, PRDX2, PRKAR1A, RIOK3, RNASE4, RPL27A, RPL37A, RPL38, RRAGD, SBF1, SH2B2, SLC35E1, SLC39A6, SNX17, SPG21, TAOK1, TEX13B, TYMP, USP34, WDR6, XRCC2, ZNF701, and ZNF721, for detecting lung cancer in a patient.

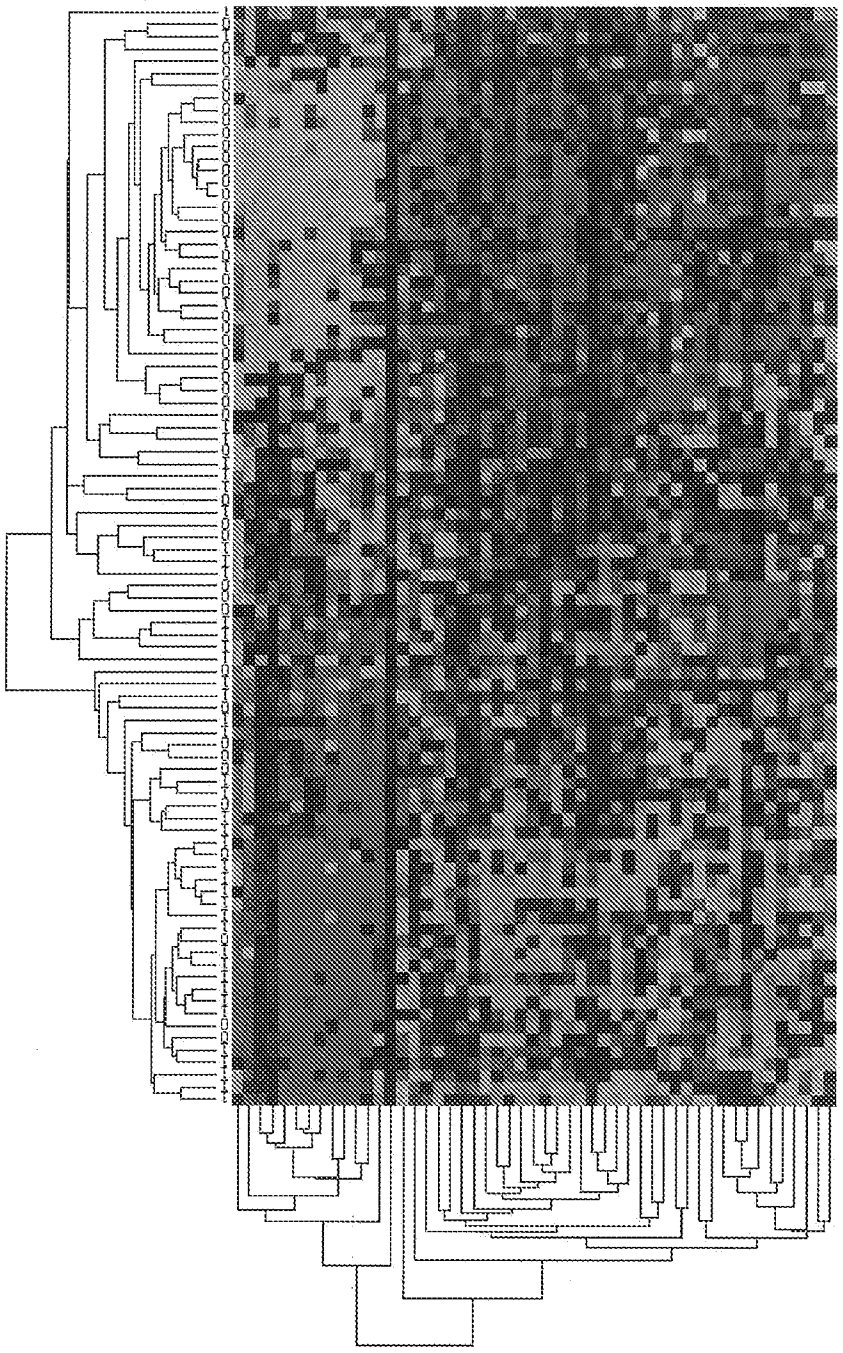
28. A method of treating lung cancer, comprising administering suitable treatment and/or prophylaxis to a patient determined to have lung cancer by the method of any one of Claims 1-25.

29. A treatment and/or prophylaxis for lung cancer in a patient determined to have lung cancer by the method of any one of Claims 1-25.

30. Use of a treatment and/or prophylaxis for lung cancer in a patient determined to have lung cancer by the method of any one of Claims 1-25

31 The method of Claim 28, treatment and/or prophylaxis of Claim 29, or use of Claim 30, wherein the treatment and/or prophylaxis is selected from the group consisting of chemotherapy, hormonal therapy, immunotherapy, radiotherapy, surgery, and combinations thereof.

FIG. 1



# FIG. 2a

## IPA Summary of Analysis

INGENUITY SYSTEMS

### Analysis settings

*View*

Reference set: Human Genome U133A
Relationship to include: Direct and Indirect
Includes Endogenous Chemicals
Optional Analyses: My Pathways My List

Filter Summary:
Consider all molecules and/or relationships

### Top Networks

| ID | Associated Network Functions | Score |
|---|---|---|
| 1 | Cell-To-Cell Signaling and Interaction, Cellular Growth and Proliferation, Gastrointestinal Disease | 28 |
| 2 | Tissue Morphology, Small Molecule Biochemistry, Cellular Development | 23 |
| 3 | Genetic Disorder, Neurological Disease, Ophthalmic Disease | 21 |
| 4 | Cellular Assembly and Organization, Cellular Movement, Cell Morphology | 10 |
| 5 | Genetic Disorder, Neurological Disease, Psychological Disorders | 3 |

(c) 2000-2011 Ingenuity Systems, Inc. All rights reserved.

INGENUITY SYSTEMS

# FIG. 2b

## Top Bio Functions

### Diseases and Disorders

| Name | p-value | # Molecules |
|---|---|---|
| Gastrointestinal Disease | 3.70E-03 - 4.35E-02 | 2 |
| Genetic Disorder | 3.70E-03 - 4.70E-02 | 6 |
| Infection Mechanism | 3.70E-03 - 1.11E-02 | 1 |
| Metabolic Disease | 3.70E-03 - 4.00E-02 | 3 |
| Cancer | 7.38E-03 - 4.70E-02 | 4 |

### Molecular and Cellular Functions

| Name | p-value | # Molecules |
|---|---|---|
| Small Molecule Biochemistry | 4.72E-04 - 1.52E-02 | 4 |
| Molecular Transport | 3.49E-03 - 3.84E-02 | 4 |
| Carbohydrate Metabolism | 3.70E-03 - 7.38E-03 | 2 |
| Cell Cycle | 3.70E-03 - 3.28E-02 | 4 |
| Cellular Assembly and Organization | 3.70E-03 - 3.28E-02 | 4 |

### Physiological System Development and Function

| Name | p-value | # Molecules |
|---|---|---|
| Cardiovascular System Development and Function | 3.70E-03 - 4.62E-02 | 3 |
| Hematological System Development and Function | 3.70E-03 - 4.70E-02 | 2 |
| Humoral Immune Response | 3.70E-03 - 3.70E-03 | 1 |
| Nervous System Development and Function | 3.70E-03 - 4.00E-02 | 3 |
| Organismal Development | 3.70E-03 - 3.64E-02 | 6 |

(c) 2000-2011 Ingenuity Systems, Inc. All rights reserved.

INGENUITY SYSTEMS

# FIG. 2c

## Top Canonical Pathways

| Name | p-value | Ratio |
|---|---|---|
| Methane Metabolism | 1.01E-05 | 3/65 (0.046) |
| Stilbene, Coumarine and Lignin Biosynthesis | 1.36E-03 | 2/74 (0.027) |
| Urea Cycle and Metabolism of Amino Groups | 4.74E-03 | 2/60 (0.025) |
| Phenylalanine Metabolism | 6.16E-03 | 2/109 (0.018) |
| Cellular Effects of Sildenafil (Viagra) | 9.53E-03 | 3/151 (0.02) |

## Top Molecules

This analysis has no expression values.

(c) 2000-2011 Ingenuity Systems, Inc. All rights reserved.

INGENUITY SYSTEMS

# FIG. 2d

**Top My Lists**

| Name | p-value | Ratio |
|---|---|---|

**Top My Pathways**

| Name | p-value | Ratio |
|---|---|---|

**Top Tox Lists**

| Name | p-value | Ratio |
|---|---|---|
| Oxidative Stress | 1.57E-02 | 2/57 (0.035) |
| Decreases Depolarization of Mitochondria and Mitochondrial Membrane | 2.92E-02 | 1/9 (0.111) |
| Hepatic Cholestasis | 7.45E-02 | 2/135 (0.015) |
| Increases Transmembrane Potential of Mitochondria and Mitochondrial Membrane | 1.54E-01 | 1/50 (0.02) |
| Renal Necrosis/Cell Death | 1.91E-01 | 1/60 (0.017) |

(c) 2000-2011 Ingenuity Systems, Inc. All rights reserved.

INGENUITY SYSTEMS

# FIG. 2e

## Top Tox Functions

### Cardiotoxicity

| Name | p-value | # Molecules |
|---|---|---|
| Cardiac Necrosis/Cell Death | 5.41E-02 - 5.41E-02 | 1 |
| Cardiac Dilation | 6.80E-02 - 6.80E-02 | 1 |
| Cardiac Proliferation | 8.52E-02 - 8.52E-02 | 1 |
| Cardiac Congestive Cardiac Failure | 2.18E-01 - 2.18E-01 | 1 |
| Heart Failure | 2.18E-01 - 2.18E-01 | 1 |

### Hepatotoxicity

| Name | p-value | # Molecules |
|---|---|---|
| Liver Hepatitis | 7.84E-02 - 1.02E-01 | 1 |
| Liver Steatohepatitis | 7.84E-02 - 7.84E-02 | 1 |
| Liver Steatosis | 7.84E-02 - 7.84E-02 | 1 |
| Hepatocellular Carcinoma | 4.68E-01 - 4.68E-01 | 1 |

### Nephrotoxicity

| Name | p-value | # Molecules |
|---|---|---|
| Renal Damage | 3.70E-03 - 3.70E-03 | 1 |
| Renal Tubule Injury | 3.70E-03 - 3.70E-03 | 1 |
| Renal Necrosis/Cell Death | 2.56E-02 - 2.56E-02 | 1 |
| Nephrosis | 1.41E-01 - 1.41E-01 | 1 |
| Kidney Failure | 2.41E-01 - 2.41E-01 | 1 |

(c) 2000-2011 Ingenuity Systems, Inc. All rights reserved.

INGENUITY SYSTEMS

FIG. 3

| | | |
|---|---|---|
| Sample size | | 90 |
| Positive group : | train 2 only = 1 | 44 |
| Negative group : | train 2 only = 0 | 46 |
| **Area under the ROC curve (AUC)** | | **0.891** |
| Standard Error [a] | | 0.0337 |
| 95% Confidence Interval [b] | | 0.808 to 0.947 |
| z statistic | | 11.617 |
| **Significance level (Area=0.5)** | **P** | **<0.0001** |

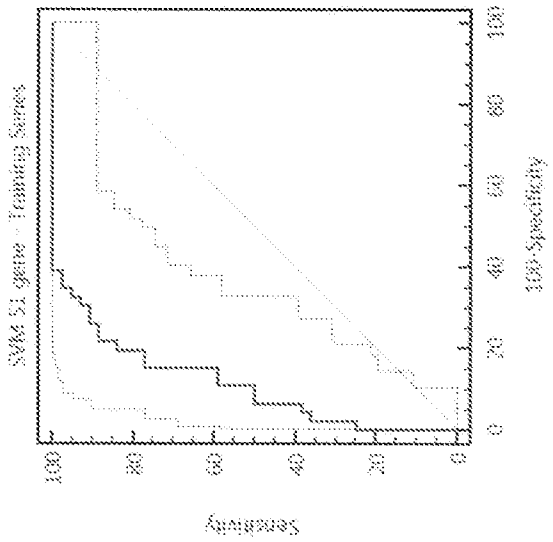


FIG. 4

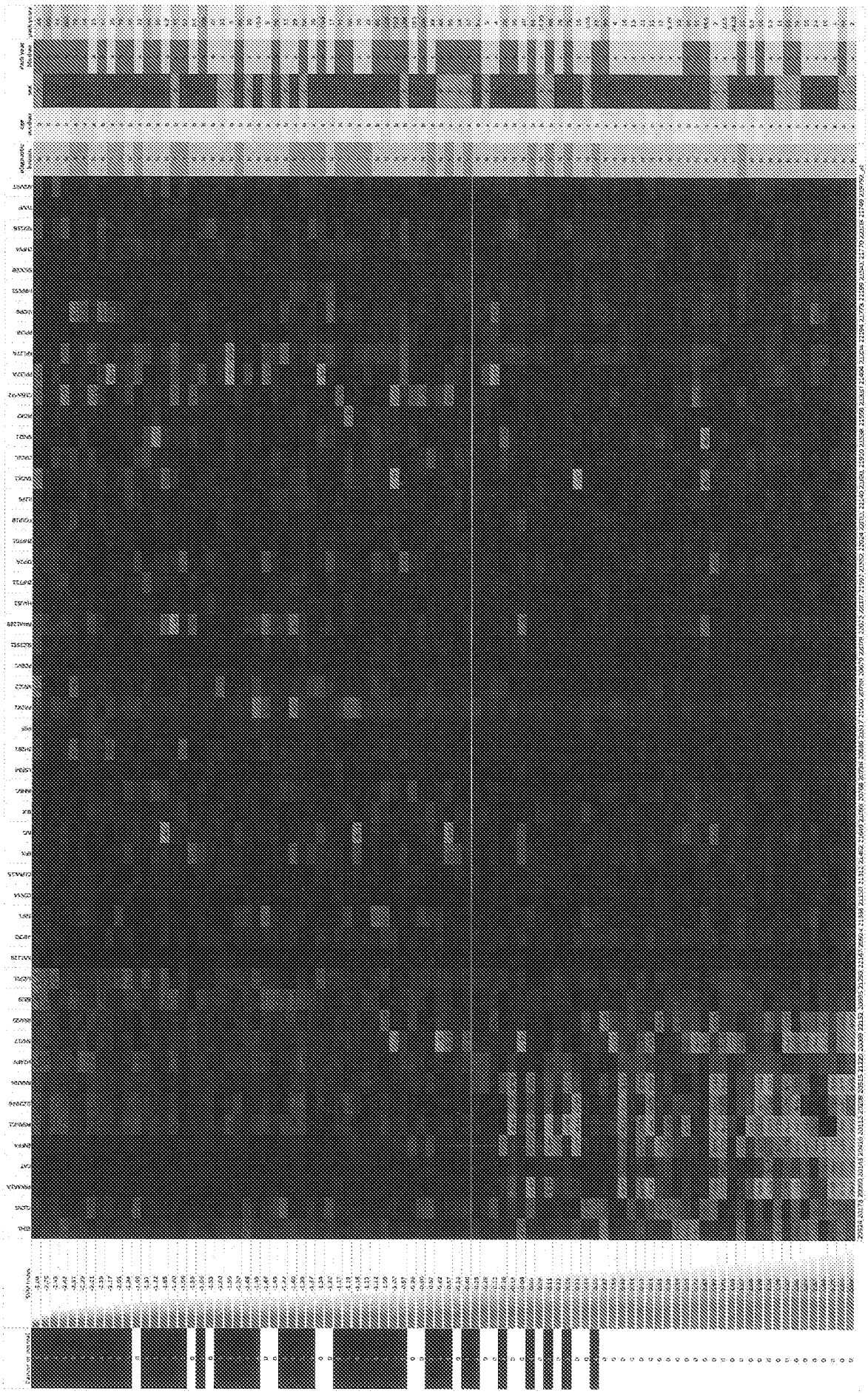


FIG. 5

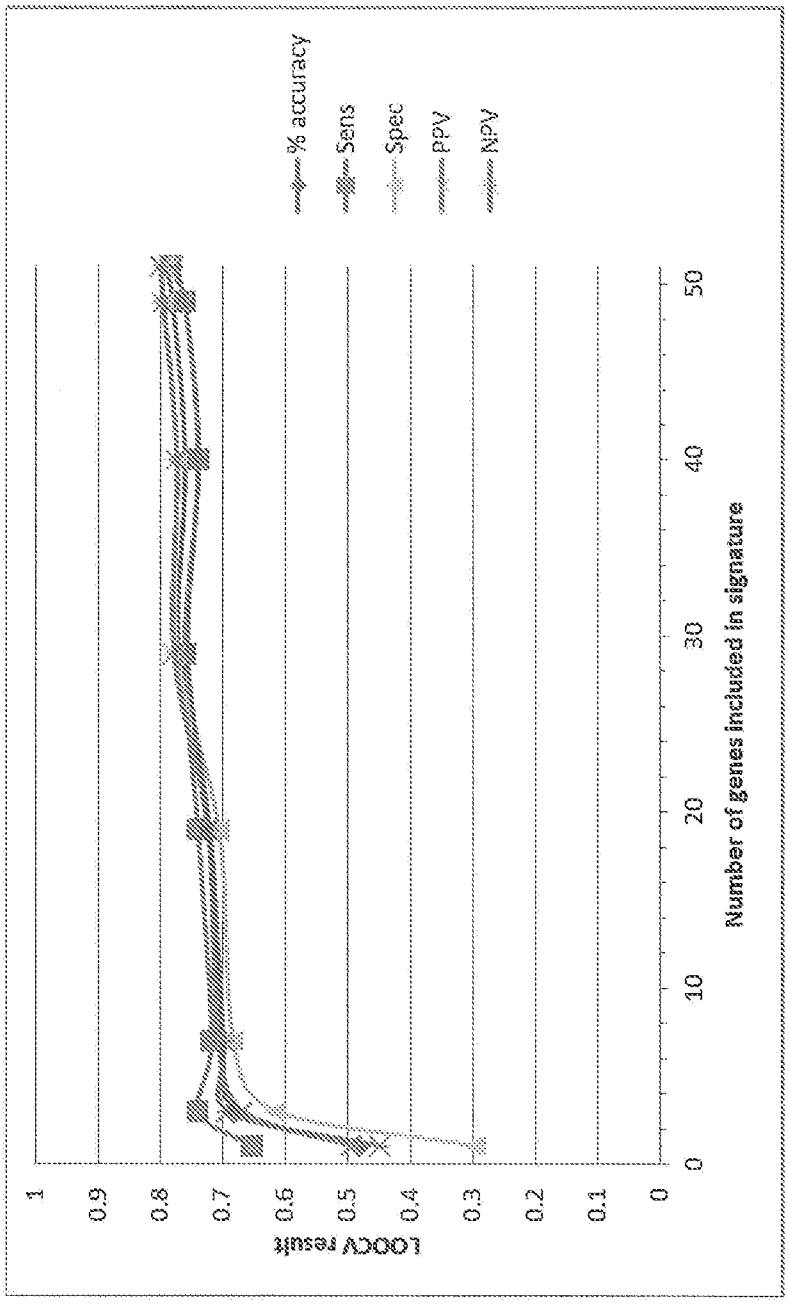

% accuracy
Sens
Spec
PPV
NPV
LOOCV result
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
10
20
30
40
50
Number of genes included in signature

FIG. 6

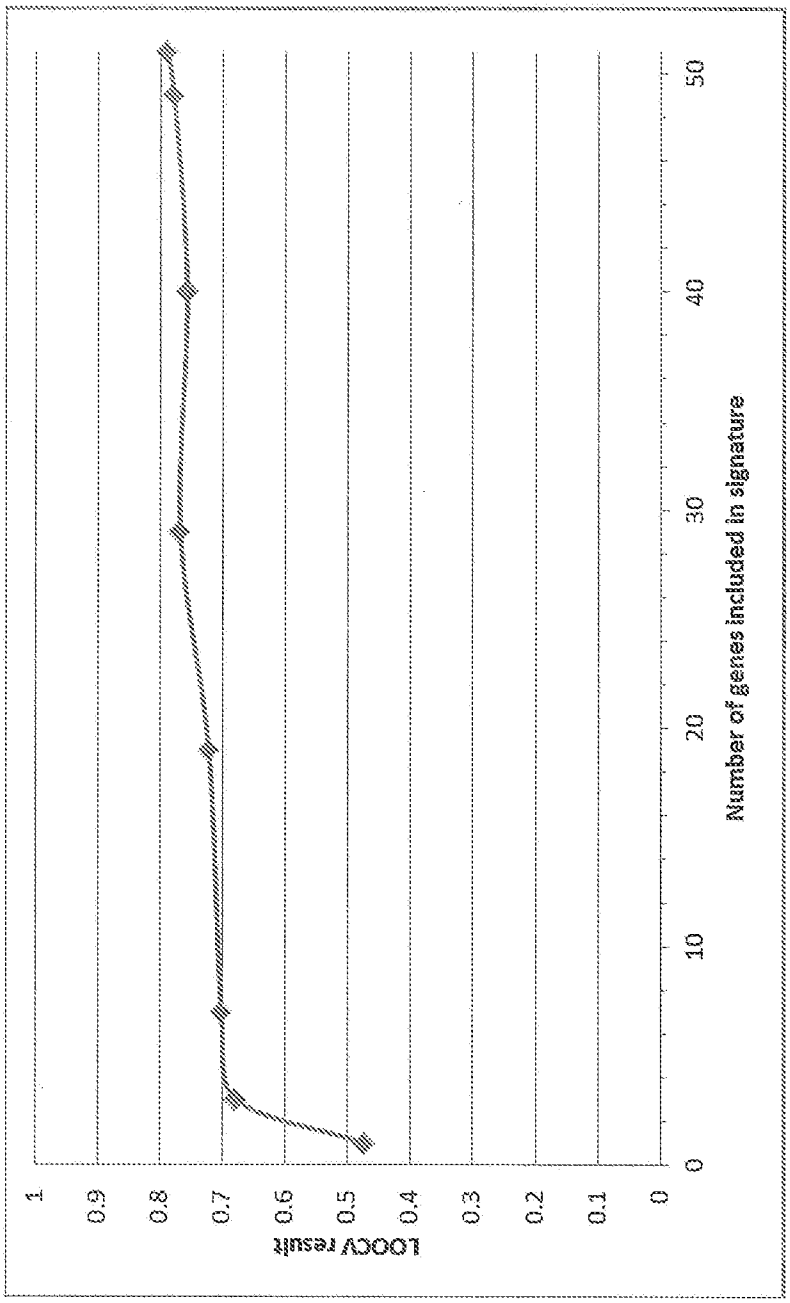

LOOCV result
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
10
20
30
40
50
Number of genes included in signature

FIG. 7

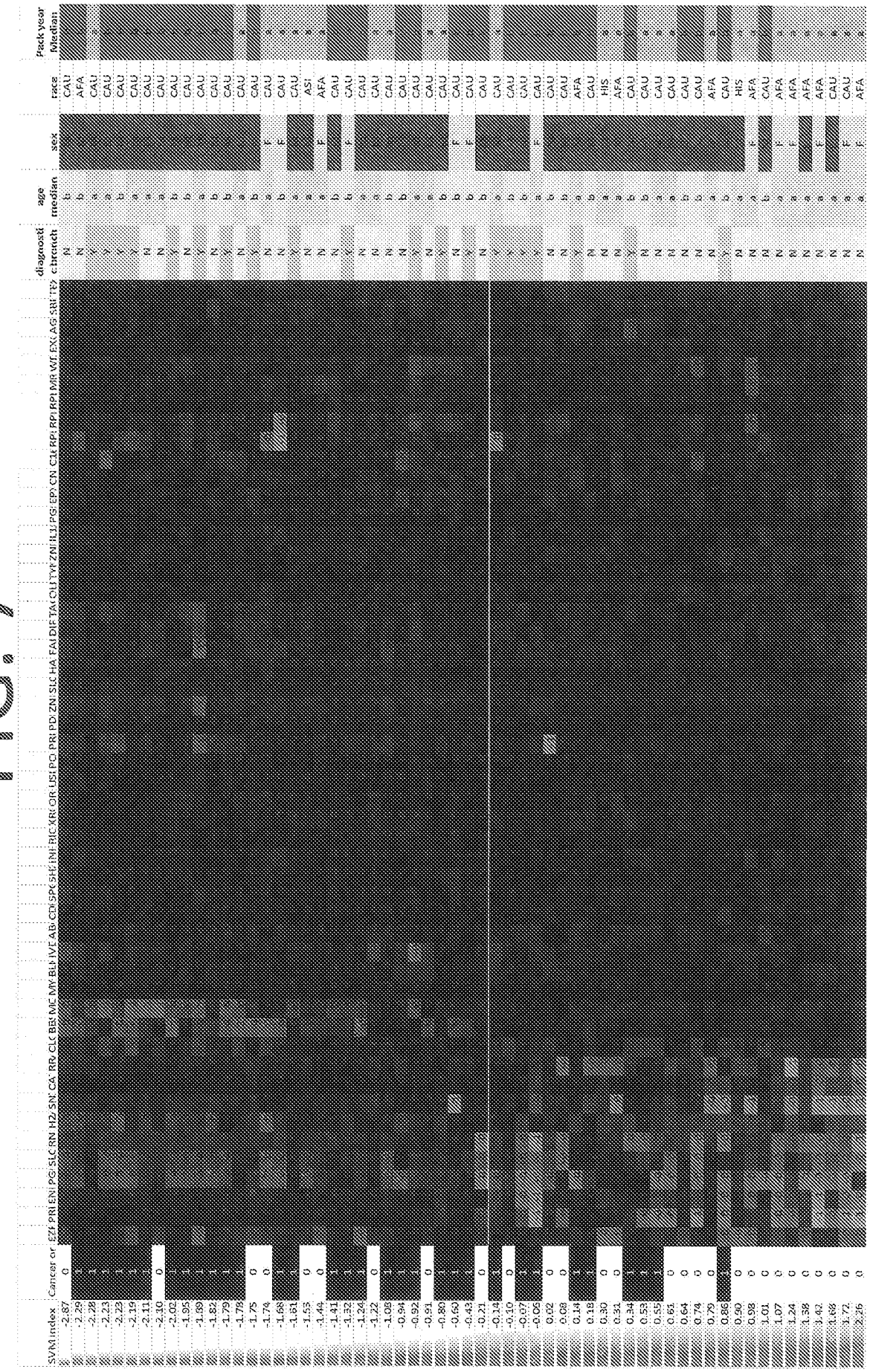

SVM Index
Cancer or
diagnosti c branch
age median
sex
race
Pack year Median

FIG. 8

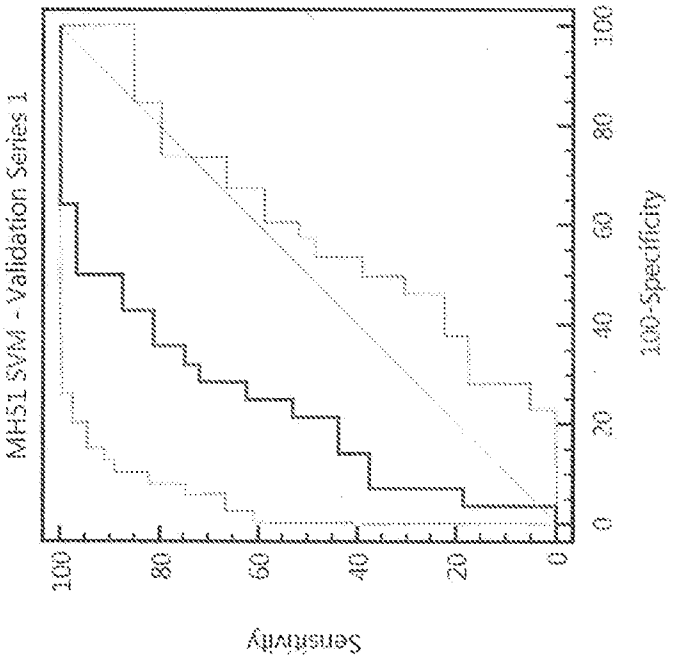

MHS1 SVM - Validation Series 1
Sensitivity
100
80
60
40
20
0
0
20
40
60
80
100
100-Specificity

FIG. 9

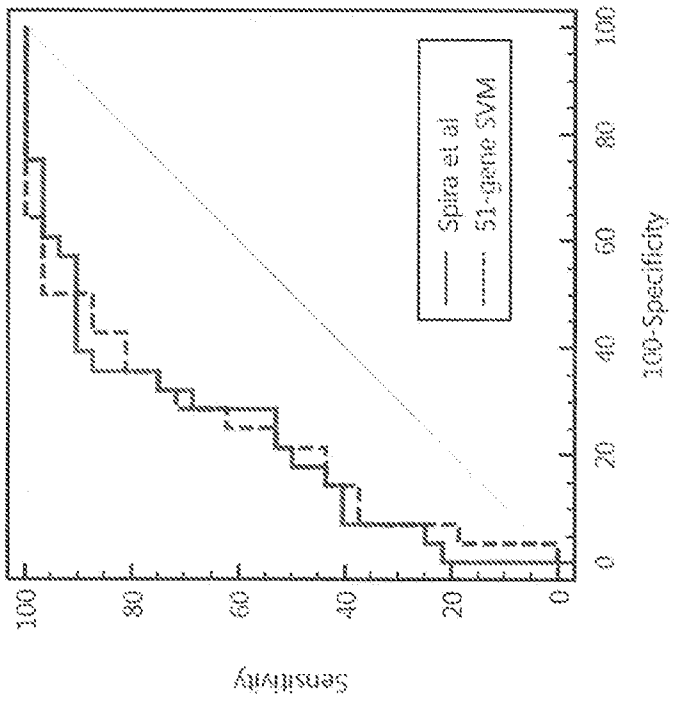

Sensitivity
100
80
60
40
20
0
0
20
40
60
80
100
100-Specificity
Spira et al
51-gene SVM

FIG. 10

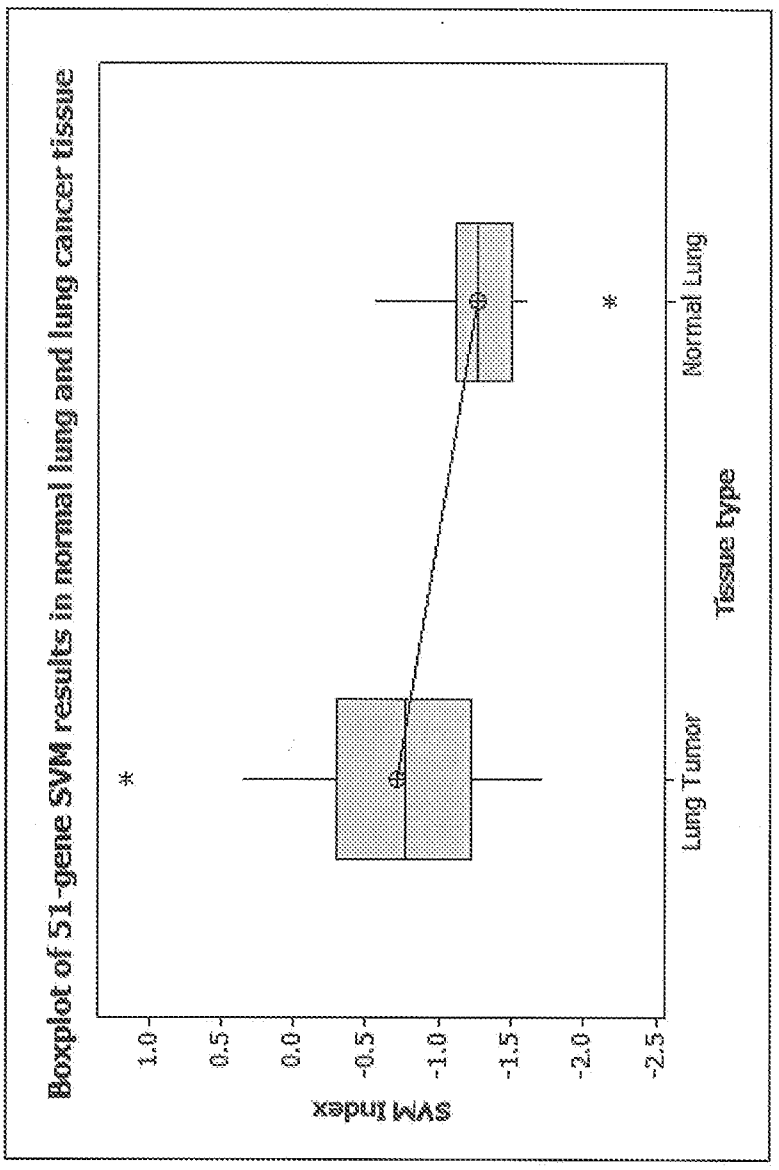

Boxplot of 51-gene SVM results in normal lung and lung cancer tissue
SVM Index
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
Lung Tumor
Normal Lung
Tissue type

FIG. 11

| | | |
|---|---|---|
| Sample size | | 41 |
| Positive group : | Cancer = 1 | 25 |
| Negative group : | Normal = 0 | 16 |

| | |
|---|---|
| Area under the ROC curve (AUC) | 0.758 |
| Standard Error [a] | 0.0747 |
| 95% Confidence Interval [b] | 0.58 to 0.877 |
| z statistic | 3.447 |
| Significance level P (Area=0.5) | 0.0006 |

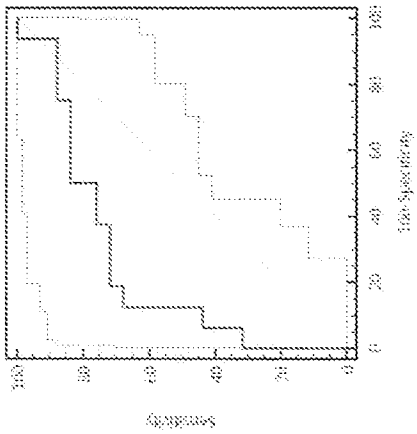


FIG. 12

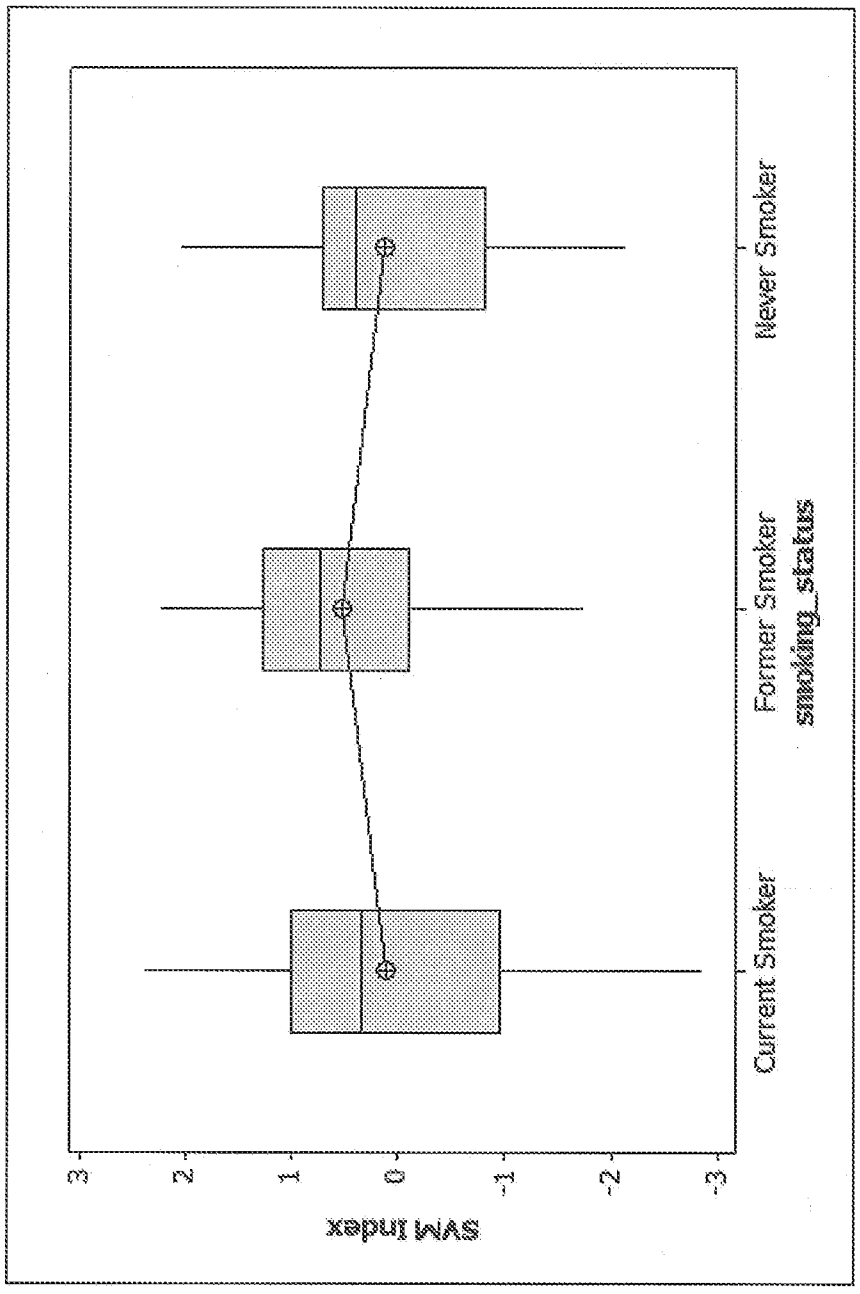
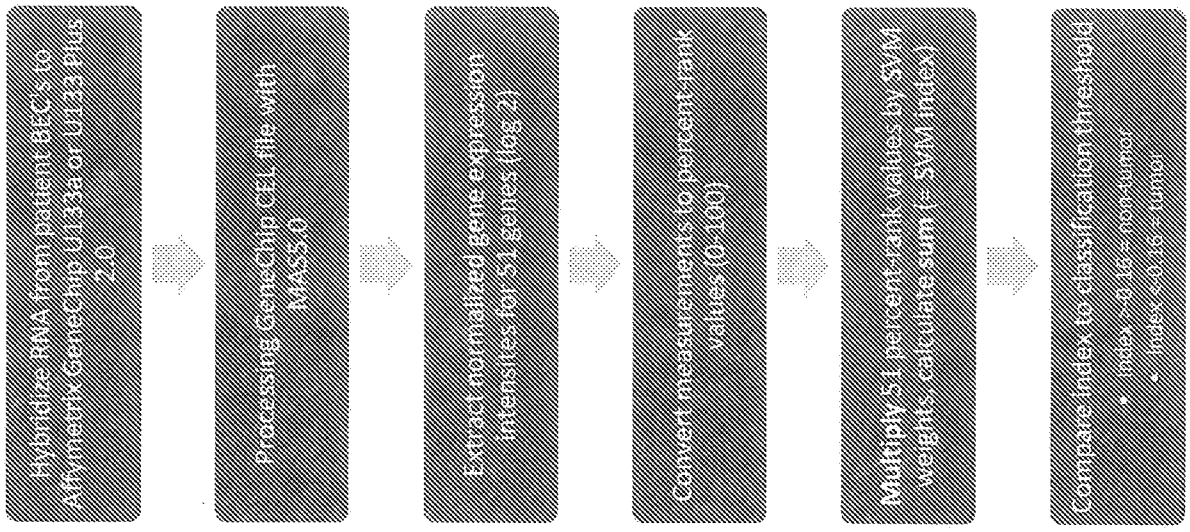

SVM Index
3
2
1
0
-1
-2
-3
Never Smoker
Former Smoker
Current Smoker
smoking_status

FIG. 13

Analysis of Variance for SVM Index, using Adjusted SS for Tests

| Source | DF | Seq SS | Adj SS | Adj MS | F | P |
|---|---|---|---|---|---|---|
| pack years | 1 | 0.000 | 1.557 | 1.557 | 1.23 | 0.271 |
| age | 1 | 1.661 | 3.675 | 3.675 | 2.90 | 0.093 |
| smoking_status | 1 | 5.148 | 5.148 | 5.148 | 4.06 | 0.047 |
| Error | 79 | 100.276 | 100.276 | 1.269 | | |
| Total | 82 | 107.085 | | | | |

FIG. 14


Hybridize RNA from patient BEC's to Affymetrix GeneChip U133a or U133 Plus 2.0
Preprocess GeneChip CEL files with MAS5.0
Extract normalized gene expression intensities for 54 selected genes
Convert measurements to percent rank values (0-100)
Multiple SVM's trained using 1 vs. 1 SVM input vectors
Compare Index to classification threshold
Index > 1.5 = Tumor
Index < 1.5 = Normal