Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer

Avrum Spira, Jennifer E Beane, Vishal Shah, Katrina Steiling, Gang Liu, Frank Schembri, Sean Gilman, Yves-Martine Dumas, Paul Calner, Paola Sebastiani, Sriram Sridhar, John Beamis, Carla Lamb, Timothy Anderson, Norman Gerry, Joseph Keane, Marc E Lenburg & Jerome S Brody

Nature Medicine, 13, 361-366, 2007

Speaker: Yi-Chiung Hsu

Dec 05, 2007

Institute of Statistical Science Academia Sinica

Dreams come true!

6/49 大樂透

★ 預估頭獎金額

$100,989,547

96/12/4

第096000097期

開出順序：

01 43 16 38 44 31

大小順序：

01 16 31 38 43 44

特別號：42 派彩結果

頭彩商店

期別	遊戲名稱	頭彩商店	地址
096000097	6/49大樂透	一直旺彩券行	高雄市苓雅區建國一路129之1號1樓
096000097	6/49大樂透	大興金香鋪	高雄市苓雅區成功一路208號
096000097	6/49大樂透	日月光彩券行	嘉義市西區湖內里民生南路５８２號１樓
096000097	6/49大樂透	雙享樂投注站	高雄市鼓山區裕誠路1139號
096000097	6/49大樂透	發利彩券行	台北市大同區南京西路３６０號１樓

獎金分配表

本期總獎金（含累積） 新台幣 1,798,524,187 元整

本期銷售總額 新台幣 1,664,370,950 元整

本期獎金分配				
獎 項	本期各獎項獎金總額 (新台幣元)	上期累積獎金總額 (新台幣元)	中獎注數 (注)	每注可得獎金 (新台幣元)
頭獎	491,688,164	783,120,165	5	254,961,665
貳獎	60,779,197	0	10	6,077,919
參獎	60,779,197	0	543	111,932
肆獎	33,766,221	0	1,262	26,756
伍獎	128,311,640	0	27,750	4,623
陸獎	35,626,000	0	35,626	1,000
普獎	204,453,600	0	511,134	400

Jerome S. Brody, M.D.
Professor of Medicine
Director of Boston University Medical
Pulmonary Pulmonary Center

Special research:

1. Genomics of smoking-related lung diseases
2. Lung cancer diagnostic/prognostic tools
3. Genomics and the pathogenesis of COPD
4. Developmental biology of the lung
5. Relation of lung cancer to lung development

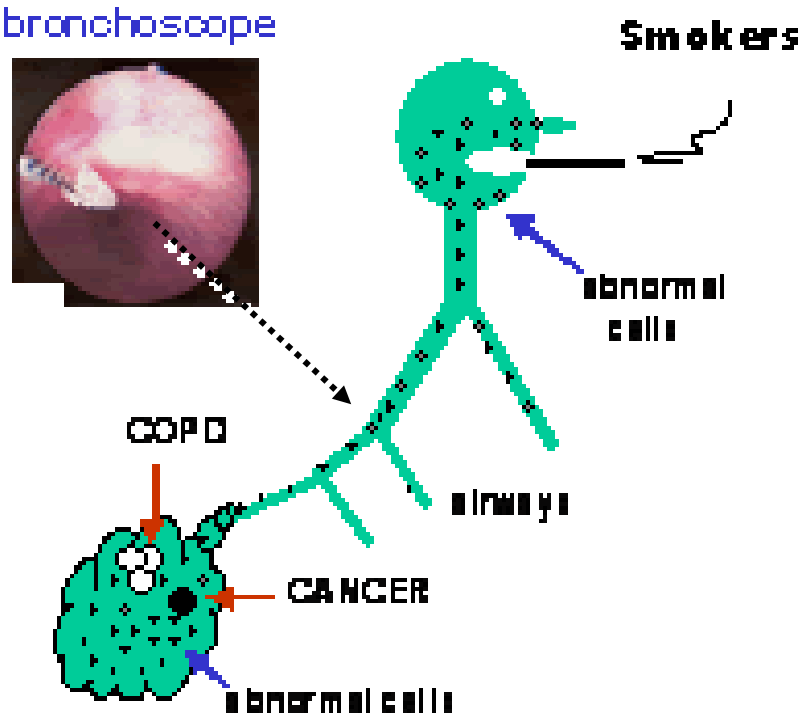
- COPD: Chronic Obstructive Pulmonary Disease

The airway epithelial cells injury in smokers

Brush airway epithelial cells

Extract RNA

Affymetrix U133A analysis



Cigarette smoking is the major cause of the two most costly and lethal pulmonary diseases, lung cancer and COPD

only 10-20% of smokers develop lung cancer; an equal number develop COPD.

PNAS. 101:10143-10148,2004..

lung cancer

- High mortality rate is related to low cure rate.
- Low cure rate is related to lack of early detection measures
- Early diagnosis is difficult because many of the symptoms of lung cancer resemble those of COPD.
- only 1%-2% of COPD patients will go on to develop lung cancer.

Histological types of lung cancer

- small-cell lung cancer

- nonsmall-cell lung cancer:
  - squamous cell carcinoma
  - adenocarcinoma
  - large-cell carcinoma

Diagnostic tests

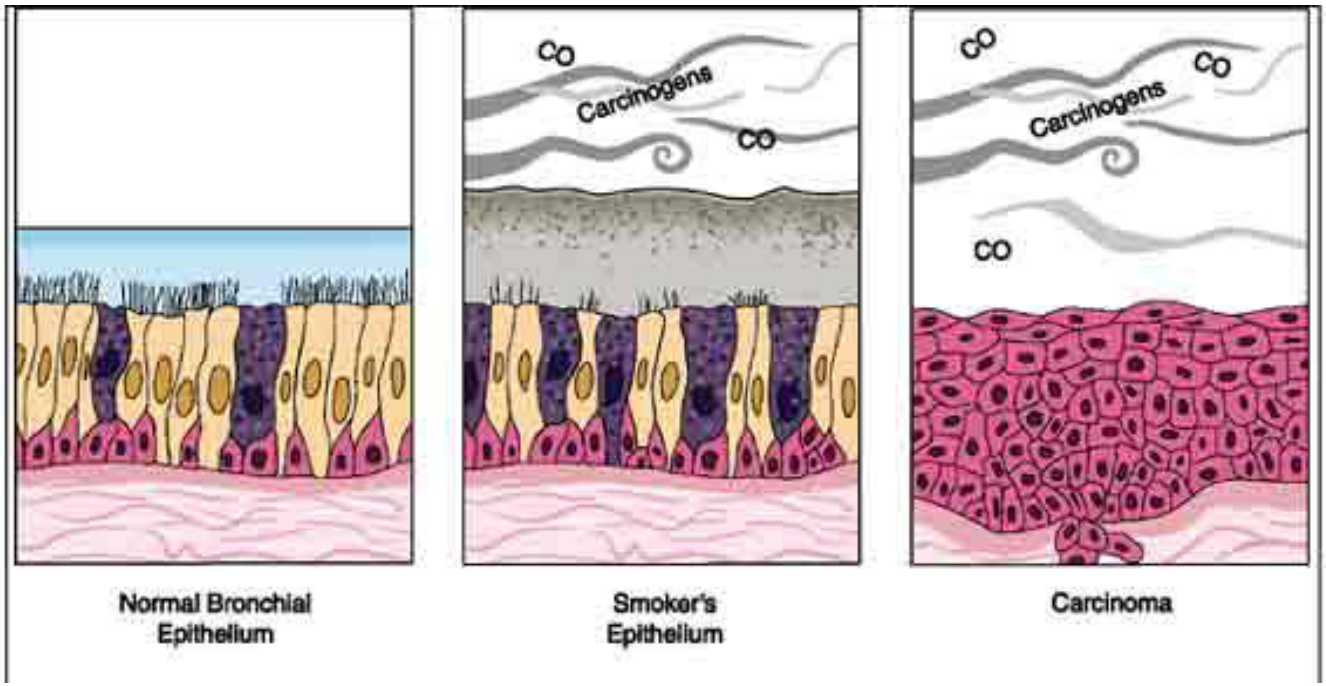
- Chest x-ray — the first step in the investigation.
- CT (computed tomography) scans — this will provide further information tumor has spread.
- Sputum analysis — sputum (from the respiratory tract).
- Needle biopsy — for cancers located closer to the ribs
- Bronchoscopy — for tumors in the main bronchi (air passages)
- A blood test may reveal certain substances, which are produced by a cancer tumor.
- Lymph nodes can be tested for cancer cells.

Problems

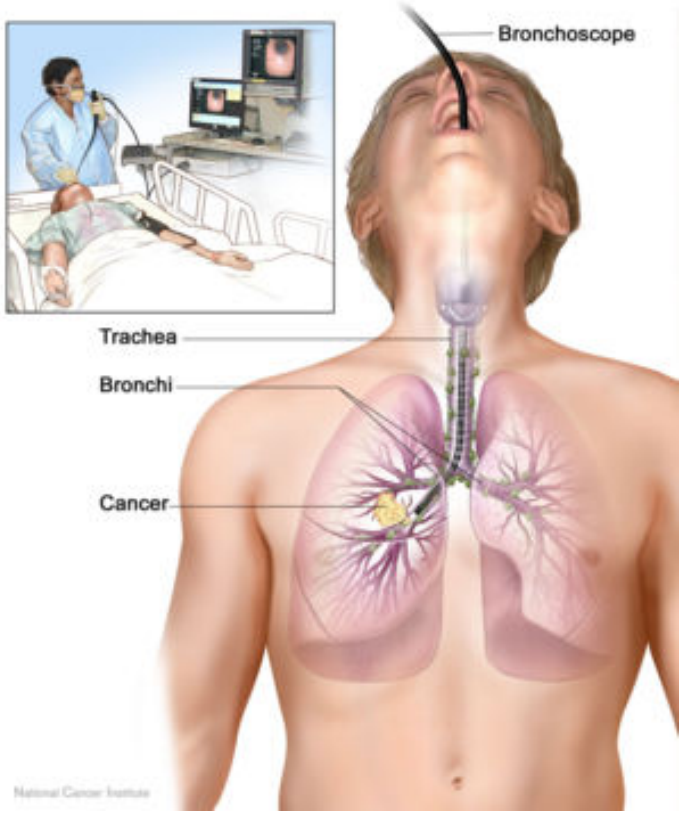
- Past screening measures: annual chest x-rays, quarterly sputum cytology have not been sensitive for lung cancer.
- Genetic features offer new possibilities

Aims

1) To distinguish smokers develop between COPD and lung cancer.
2) To develop new approaches for the early diagnosis of lung cancer and for assessing a smokers' risk of developing lung cancer.



Bronchoscopy



1. This is a test that looks at the inside of the airways.
2. A flexible tube called a bronchoscope is put into the airway.
3. The tube has an eyepiece so that the doctor can see into your airways.
4. Biopsies (samples of tissue and cells) can also be taken during a bronchoscopy.
   These are sent to a laboratory for testing to see if there are any cancer cells present.

Approaches (I)

Study patients and sample collection

- **A. Primary sample set**
- **B. Prospective sample set**
- **C. Airway epithelial cell collection**

Primary sample set

- Samples collection between Jan 2003 and Apr 2005 obtains former and current smokers
- >21 yr, no contraindications to bronchoscopy
- n=152

Prospective sample set

- Samples collection between May 2005 and Dec
- 2005 obtains former and current smokers
- >21 yr, no contraindications to bronchoscopy
- n=40

Microarray data acquisition and preprocessing

- **A. Microarray data acquisition**
- **B. Preprocessing of array data**
- **C. Sample filter**
- **D. Prospective validation test set**

- Affymetrix HG-U133A array
- Normalization (RMA)
- Sample filter (MAS 5.0 + average z-score)
- An average z-score > 0.129 was excluded
- Primary sample set: n=129
- Prospective sample set: n=35

Supplementary Table 1. Patient demographics by dataset and cancer status.

	(n=129) Primary Dataset		(n=35) Prospective Dataset	
	Cancer	Non Cancer	Cancer	Non Cancer
Samples	60	69	18	17
Age (years)	64.1 ± 9.0*	49.8 ± 15.2*	66.1 ± 11.4	62.2 ± 11.1
Smoking Status	51.7% F, 48.3% C	37.7% F, 62.3% C	66.7% F, 33.3% C	52.9% F, 47.1% C
Gender	80% M, 20% F	73.9% M, 26.1% F	66.7% M, 33.3% F	70.6% M, 29.4% F
Pack-Years	57.4 ± 25.6*	29.4 ± 27.3*	46.7 ± 28.8	60 ± 44.3
Age Started (years)	15.2 ± 4.2	16.7 ± 6.8	16.4 ± 7.3	14.2 ± 3.8
Smoking Intensity (PPD): Currents	1.3 ± 0.45*	0.9 ± 0.5*	1.1 ± 0.44	1.2 ± 0.9
Months Quit: Former	113 ± 118	158 ± 159	153 ± 135	93 ± 147

PPD = packs per day, Smoking status: *F* = former smokers, *C* = current smokers, Gender: *M* = male, *F* = female. (Mean ± SD) 1 pack has 20 cigarettes

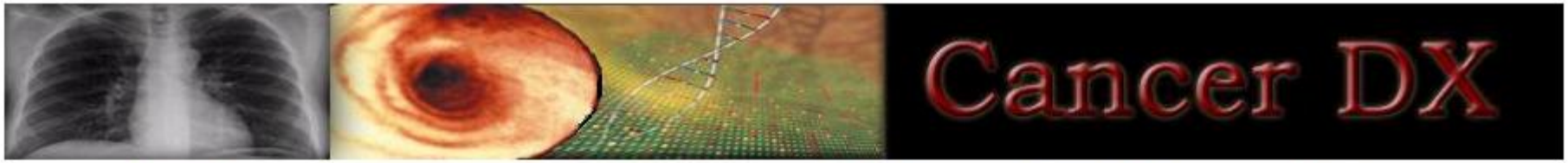
smoking history: pack-years = (number of cigarettes smoked per day x number of years smoked)/20

Supplementary Table 2. Cell type and staging information for the 60 lung cancer patients in the *n* = 129 primary dataset.

	Cancer Samples
Cell Type:	
Non Small Cell	48
- Squamous Cell Carcinoma	23
- Adenocarcinoma (BAC)	11 (1)
- Large Cell Carcinoma	4
- Unclassified	10
Small Cell	11
Unknown	1
Stage:	
Stage Ia	2
Stage Ib	9
Stage IIa	2
Stage IIb	0
Stage IIIa	9
Stage IIIb	9
Stage IV	17

Detailed pathological descriptions of cancer specimens are available at http://pulm.bumc.bu.edu/CancerDx/

➢BCA: Bronchoalveolar carcinoma



Introduction:

Lung cancer is the leading cause of cancer death in both men and women in the United States. Eighty-five to ninety percent of subjects with lung cancer are current or former smokers, yet only 10-20% of heavy smokers actually develop lung cancer. The high mortality in patients with lung cancer (80-85% in five years) results from our inability to identify which of the 90 million current and former smokers in the United States are at greatest risk for developing lung cancer and from the lack of effective tools to diagnose the disease at an early stage before it has spread to regional nodes or has metastasized beyond the lung.

Based on the concept that cigarette smoking creates a respiratory tract "field defect" , we have identified a profile of gene expression in relatively easily accessible large airway epithelial cells obtained at bronchoscopy that can serve as an indicator of the amount and type of cellular injury induced by smoking and provide a diagnostic tool in smokers who are being evaluated for the possibility of lung cancer. Histologically normal bronchial airway epithelial cells were collected from brushings of the right mainstem bronchus in current and former smoking subjects (n=152) undergoing fiberoptic bronchoscopy for clinical suspicion of lung cancer. RNA was extracted and hybridized to the Affymetrix HG-U133A microarray(containing ~22,500 transcripts). Patients were followed until final diagnosis of lung cancer or an alternate benign diagnosis was made. Following data preprocessing, 129 samples were divided into a training set (n=77) and independent test set (n=52). A weighted voting algorithm was used to build a predictive committee of 80 genes that could distinguish the 2 classes in the training set. This 80 gene predictor represents a biomarker that was found to be a highly sensitive (80%) and specific (84%) diagnostic for lung cancer when

http://pulm.bumc.bu.edu/CancerDx/

Approaches (II)

Microarray data analysis

A. Class prediction algorithm

B. Randomization

C. Characteristics of the 1000 additional runs of the algorithm

D. Comparison of RMA vs. MAS5.0 and weighted voting vs. PAM

E. Cancer cell type and stage

- Weighted voting algorithm
- Adjustment: ANCOVA with pack-years as the covariate.
- Gene selection: signal to noise metric and internal cross-validation 40 up 40 down probesets

Supplementary Table 1. Patient demographics by dataset and cancer status.

	(n=129) Primary Dataset		(n=35) Prospective Dataset	
	Cancer	Non Cancer	Cancer	Non Cancer
Samples	60	69	18	17
Age (years)	64.1 ± 9.0*	49.8 ± 15.2*	66.1 ± 11.4	62.2 ± 11.1
Smoking Status	51.7% F, 48.3% C	37.7% F, 62.3% C	66.7% F, 33.3% C	52.9% F, 47.1% C
Gender	80% M, 20% F	73.9% M, 26.1% F	66.7% M, 33.3% F	70.6% M, 29.4% F
Pack-Years	57.4 ± 25.6*	29.4 ± 27.3*	46.7 ± 28.8	60 ± 44.3
Age Started (years)	15.2 ± 4.2	16.7 ± 6.8	16.4 ± 7.3	14.2 ± 3.8
Smoking Intensity (PPD): Currents	1.3 ± 0.45*	0.9 ± 0.5*	1.1 ± 0.44	1.2 ± 0.9
Months Quit: Former	113 ± 118	158 ± 159	153 ± 135	93 ± 147

PPD = packs per day, Smoking status: *F* = former smokers, *C* = current smokers, Gender: *M* = male, *F* = female. (Mean ± SD) 1 pack has 20 cigarettes

smoking history: pack-years = (number of cigarettes smoked per day x number of years smoked)/20

Approaches (II)

Microarray data analysis

- **A. Class prediction algorithm**
- **B. Randomization**
- **C. Characteristics of the 1000 additional runs of the algorithm**
- **D. Comparison of RMA vs. MAS5.0 and weighted voting vs. PAM**
- **E. Cancer cell type and stage**

Fig. 1a. Class prediction methodology

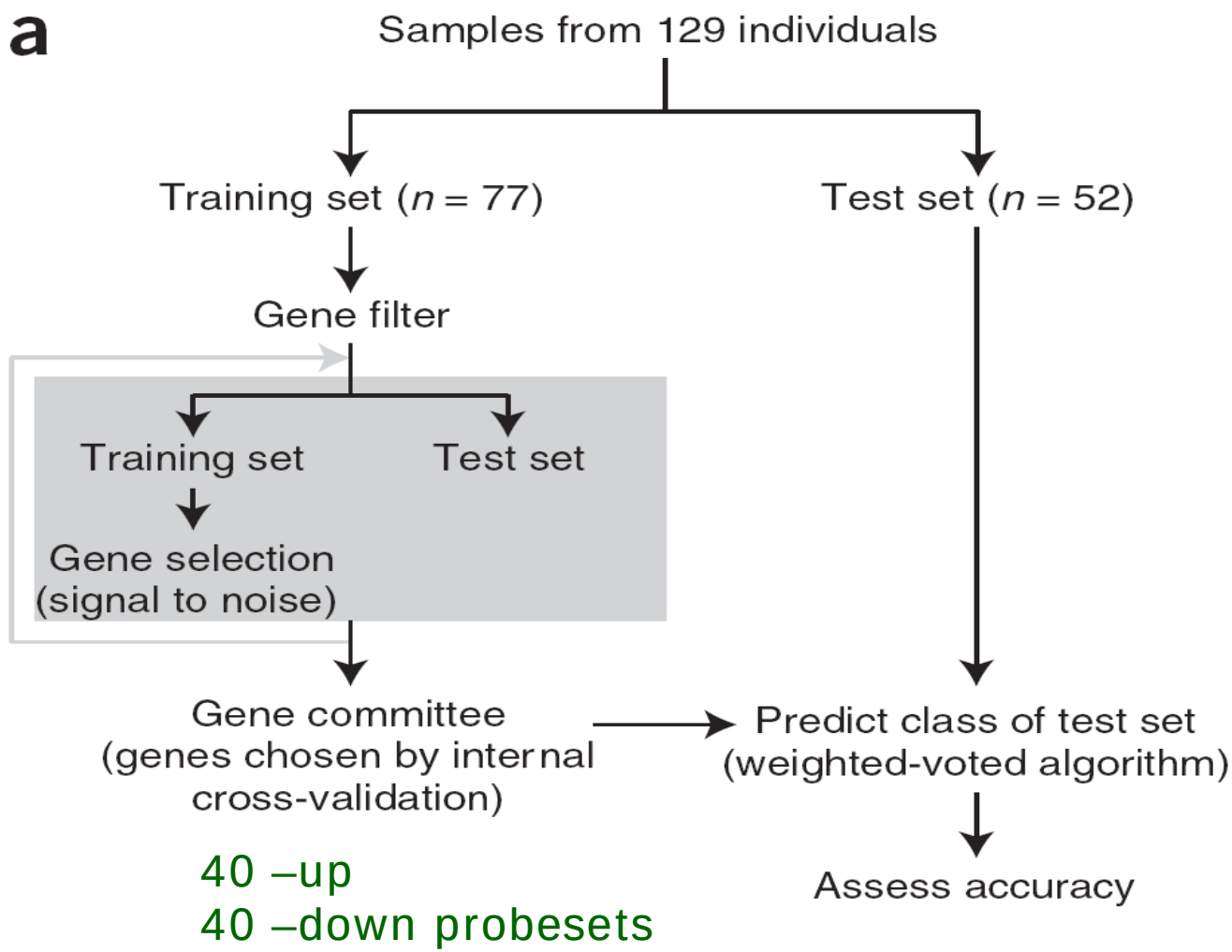
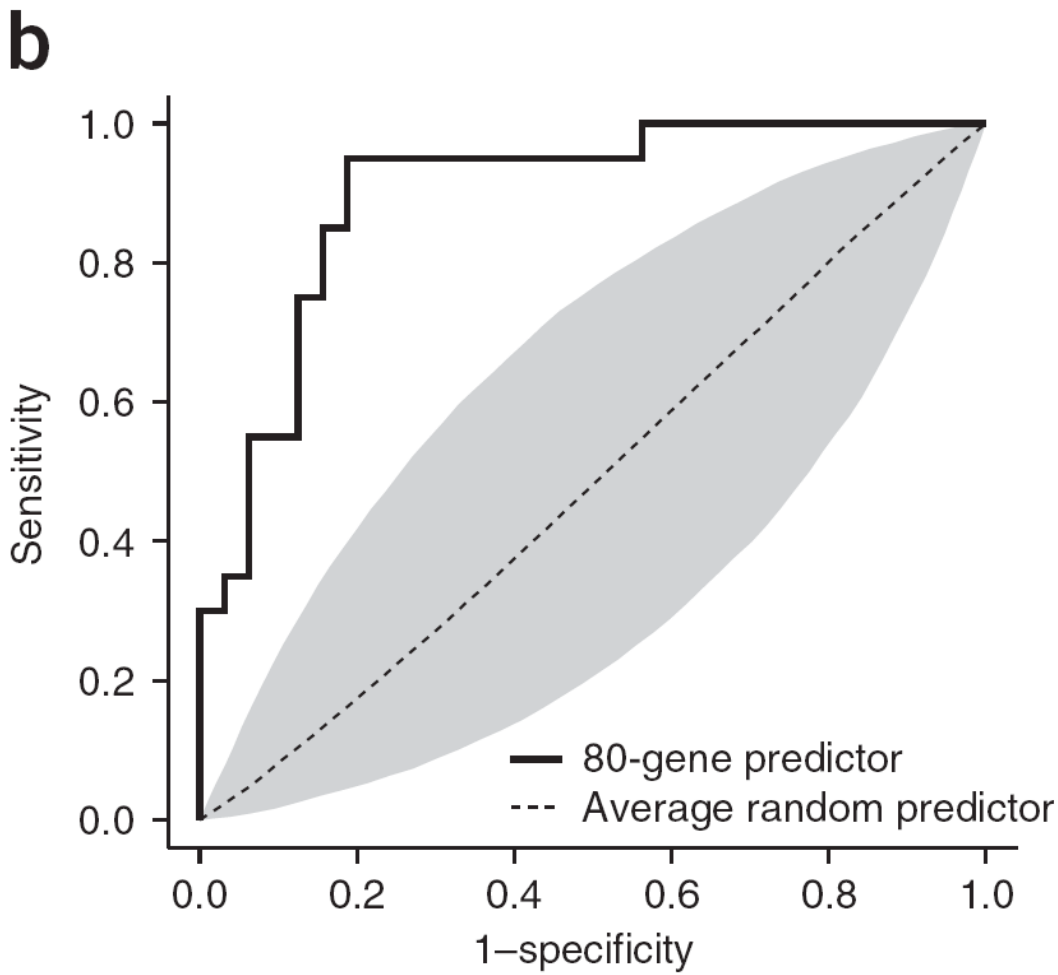


Fig. 1b. Biomarker performance.

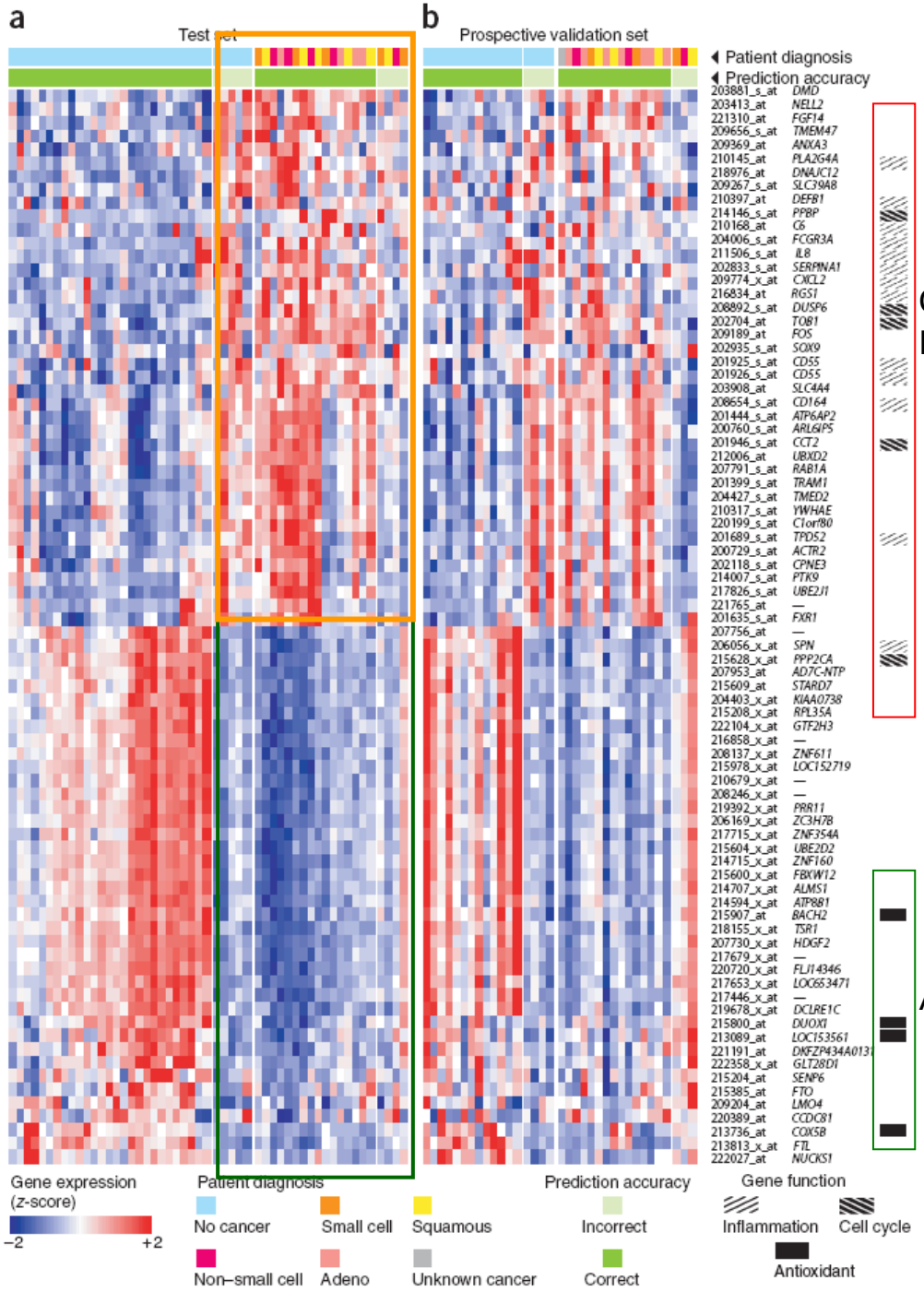


P=0.004

Fig. 2. Hierarchical clustering of 80 probeset expression in two sets.

2a. Test set (n=52)
accuracy 83%
sensitivity 80%
specificity 84%

2b. Prospective set (n=35)
accuracy 80%
sensitivity 83%
specificity 76%



Supplementary Table 4. Comparing the airway biomarker to randomized biomarkers.

	Accuracy			Sensitivity			Specificity			AUC		
	Mean	*SD*	*P*	*Mean*	*SD*	*P*	*Mean*	*SD*	*P*	*Mean*	*SD*	*P*
Actual Classifier	0.827			0.800			0.844			0.897		
Random 1	0.491	0.171	0.018	0.487	0.219	0.114	0.493	0.185	0.015	0.487	0.223	0.004
Random 2	0.495	0.252	0.078	0.496	0.249	0.173	0.495	0.263	0.073	0.495	0.309	0.008
Random 3	0.495	0.193	0.021	0.491	0.268	0.217	0.498	0.170	0.006	0.492	0.264	0.007
1000 Runs	0.784	0.054	0.283	0.719	0.104	0.245	0.830	0.060	0.407	0.836	0.053	0.108
1000 Runs Random 1	0.504	0.126	0.002	0.501	0.154	0.025	0.506	0.154	0.003	0.507	0.157	0.001

Random 1: training set were permuted and the entire algorithm, including gene selection, was re-run

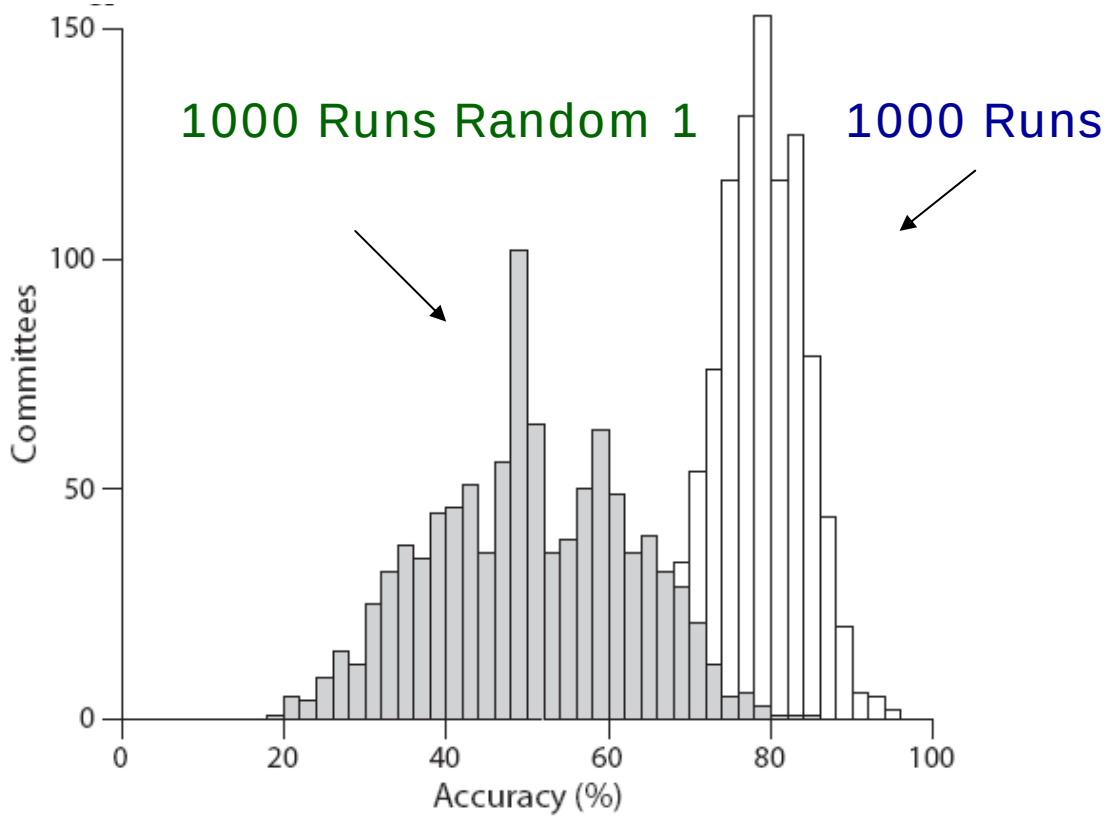
Random 2: the 80 genes in the original predictor permuted the class labels of the training set samples

Random 3: selecting 80 random probesets where the class labels of the training set were retained

1000 Runs: primary data set using the same algorithm as was used to build the actual biomarker to create 1,000 additional training

- The *P*-values indicate the percentage of 1000 runs that had the same performance than the actual classifier.

Supplementary Fig. 3 Biomarker accuracy is independent of the composition of the training set.



Approaches (II)

Microarray data analysis

- **A. Class prediction algorithm**
- **B. Randomization**
- **C. Characteristics of the 1000 additional runs of the algorithm**
- **D. Comparison of RMA vs. MAS5.0 and weighted voting vs. PAM**
- **E. Cancer cell type and stage**

- The different class-prediction and data preprocessing algorithms were identical.
- Samples did not separate by cell type or stage in this analysis (PCA)

Approaches (III)

Validation of differential expression of select biomarker genes

A. QPCR

B. Immunohistochemistry

Link to lung cancer tissue microarray datasets

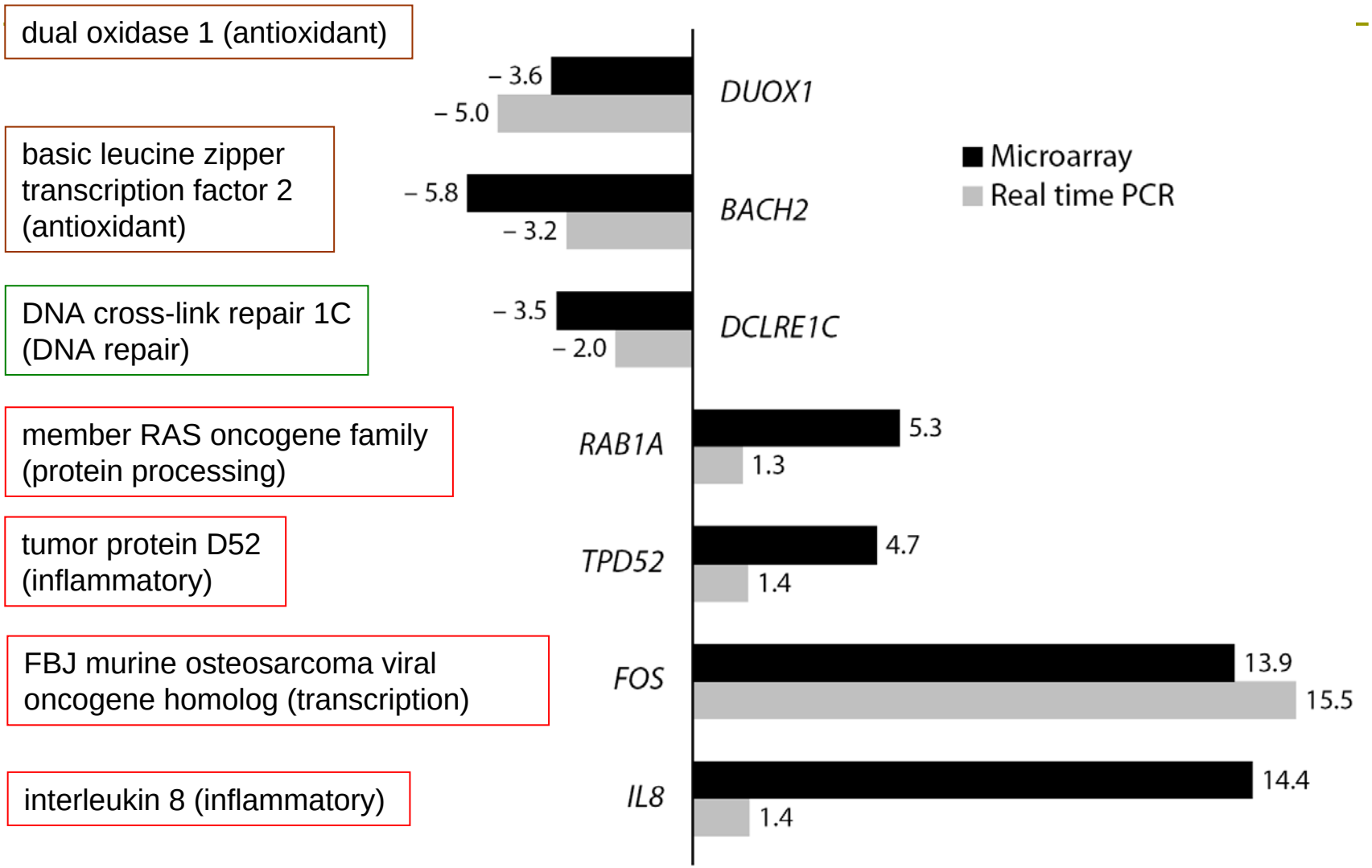
A. Analyses of Bhattacharjee dataset

B. Analyses of Wachi dataset

C. Analyses of Raponi dataset

D. Analyses of Potti dataset

Supplementary Fig. 1 Confirmation of expression differences for selected biomarker genes by RT-PCR.



Supplementary Table 3. Functional classification of biomarker genes. (I)

Up-regulated		Down-regulated	
Gene Symbol	Weight	Gene Symbol	Weight
Inflammation/Immune Function			
IL8	0.31	*SPN*	0.2
CD55	0.26		
RGS1	0.26		
PLA2G4A	0.26		
C6	0.24		
DEFB1	0.23		
TPD52	0.22		
CD164	0.21		
CXCL2	0.21		
SERPINA1	0.21		
FCGR3A	0.2		
Cell Cycle			
TOB1	0.26	*PPP2CA*	0.16
DUSP6	0.25		
CCT2	0.25		
PPBP	0.24		
Receptor Signaling			
ATP6AP2	0.21	*NUCKS1*	0.17
PTK9	0.21	*HDGF2*	0.17
ANXA3	0.2		
FGF14	0.2		

Up-regulated		Down-regulated	
Gene Symbol	Weight	Gene Symbol	Weight
Cytoskeleton/Cell Adhesion			
DMD	0.24		
NELL2	0.24		
ACTR2	0.22		
CPNE3	0.21		
Transcription			
FOS	0.27	*ZC3H7B*	0.26
SOX9	0.26	*CCDC81*	0.24
UBXD2	0.22	*ZNF354A*	0.22
		ZNF160	0.2
		ZNF611	0.19
		LMO4	0.16

Supplementary Table 3. Functional classification of biomarker genes. (II)

Up-regulated		Down-regulated	
Gene Symbol	Weight	Gene Symbol	Weight
Protein Processing			
YWHAE	0.3	*LOC653471*	0.27
TMED2	0.24	*RPL35A*	0.23
DNAJC12	0.23	*GLT28D1*	0.19
RAB1A	0.22	*TSR1*	0.19
TRAM1	0.21		
Antioxidant			
		BACH2	0.18
		LOC153561	0.16
		COX5B	0.16
		DUOX1	0.16
Ubiquitination			
		UBE2D2	0.22
		SENP6	0.17
		FBXW12	0.17
DNA Repair			
		GTF2H3	0.19
		DCLRE1C	0.17

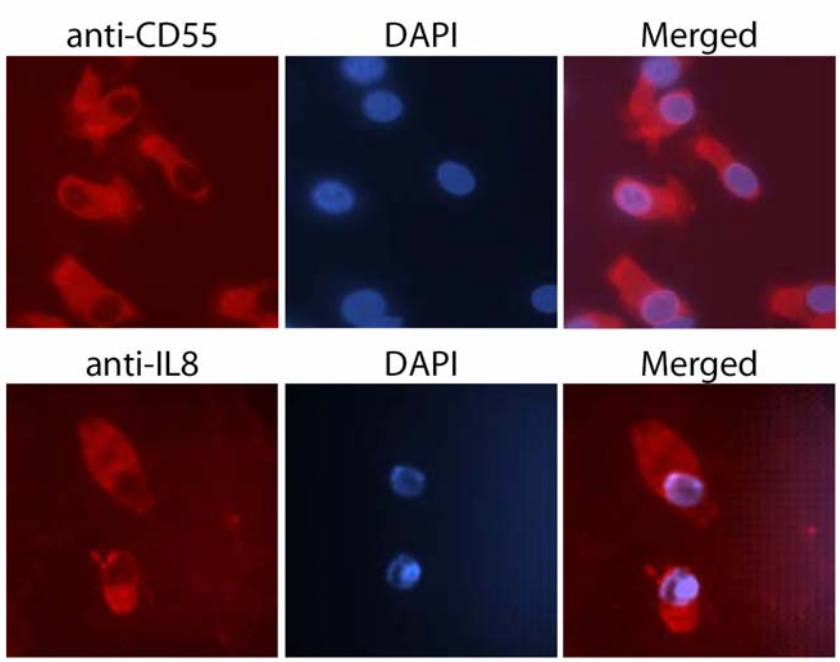
Up-regulated		Down-regulated	
Gene Symbol	Weight	Gene Symbol	Weight
Miscellaneous			
SLC39A8	0.24	*ATP8B1*	0.28
SLC4A4	0.23	*AD7C-NTP*	0.22
TMEM47	0.23	*STARD7*	0.19
UBE2J1	0.23	*FTO*	0.19
FXR1	0.22	*DKFZP434A0131*	0.19
ARL6IP5	0.21	*FTL*	0.18
C1orf80	0.21	*FLJ14346*	0.18
		PRR11	0.17
		KIAA0738	0.17
		ALMS1	0.16
		LOC152719	0.16

Supplementary Table 3. Functional classification of biomarker genes. (I)

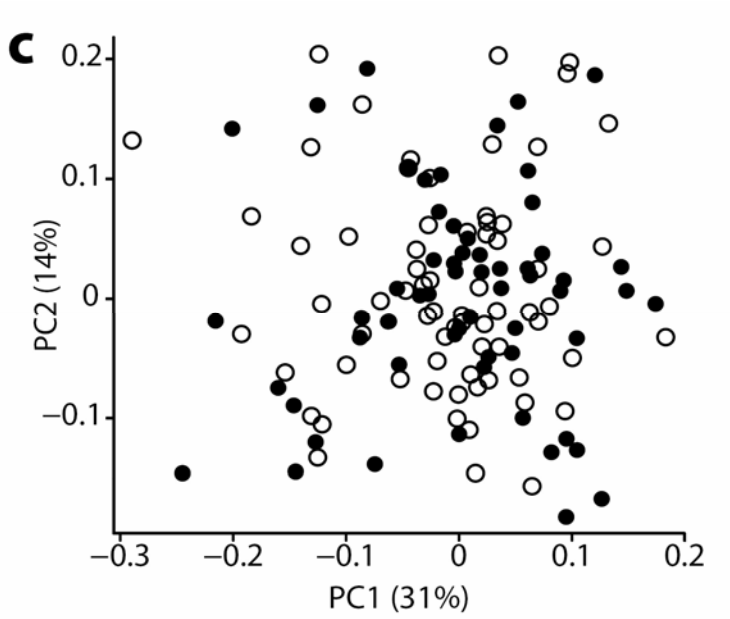
Up-regulated		Down-regulated	
Gene Symbol	Weight	Gene Symbol	Weight
Inflammation/Immune Function			
IL8	0.31	*SPN*	0.2
CD55	0.26		
RGS1	0.26		
PLA2G4A	0.26		
C6	0.24		
DEFB1	0.23		
TPD52	0.22		
CD164	0.21		
CXCL2	0.21		
SERPINA1	0.21		
FCGR3A	0.2		
Cell Cycle			
TOB1	0.26	*PPP2CA*	0.16
DUSP6	0.25		
CCT2	0.25		
PPBP	0.24		
Receptor Signaling			
ATP6AP2	0.21	*NUCKS1*	0.17
PTK9	0.21	*HDGF2*	0.17
ANXA3	0.2		
FGF14	0.2		

Up-regulated		Down-regulated	
Gene Symbol	Weight	Gene Symbol	Weight
Cytoskeleton/Cell Adhesion			
DMD	0.24		
NELL2	0.24		
ACTR2	0.22		
CPNE3	0.21		
Transcription			
FOS	0.27	*ZC3H7B*	0.26
SOX9	0.26	*CCDC81*	0.24
UBXD2	0.22	*ZNF354A*	0.22
		ZNF160	0.2
		ZNF611	0.19
		LMO4	0.16

Supplementary Fig. 2 Inflammatory gene expression in bronchial epithelial cells.



Bronchial epithelial cells



11 inflammatory gene probesets
Primary dataset (n=129)

CD55: decay accelerating factor for complement
IL-8: Interleukin-8

Link to Lung Cancer Tissue Microarray Datasets (I)

Analyses of Bhattacharjee dataset

- 128 samples for further analysis (88 adenocarcinomas, three small cell, 20 squamous, and 17 normal lung samples).
- The samples were classified with 89.8% accuracy, 89.1% sensitivity, and 100% specificity.
- The airway biomarker classified normal lung tissue from smokers without cancer and lung tumor tissue from smokers with 90% accuracy.

PNAS 98,13790–13795, 2001

Link to Lung Cancer Tissue Microarray Datasets (II)

Analyses of Wachi dataset

- The 10 arrays represent five squamous cell lung cancer tissue samples from smokers and five matched adjacent histologically normal lung tissue samples taken from the same patients.
- All samples were classified as being from smokers with cancer; moreover, the expression of biomarker probesets was similar between tumor and adjacent normal tissue samples

.

Bioinformatics 21, 4205–4208, 2005

Link to Lung Cancer Tissue Microarray Datasets (III)

Analyses of Raponi dataset

- These 130 samples represent fresh frozen, surgically resected malignant lung tissue from 129 individual patients with different stages of squamous cell carcinoma.

- 80 biomarker correctly classified 99% (129 of 130) of samples.

Cancer Res. 66, 7466–7472, 2006

Link to Lung cancer Tissue Microarray Datasets (IV)

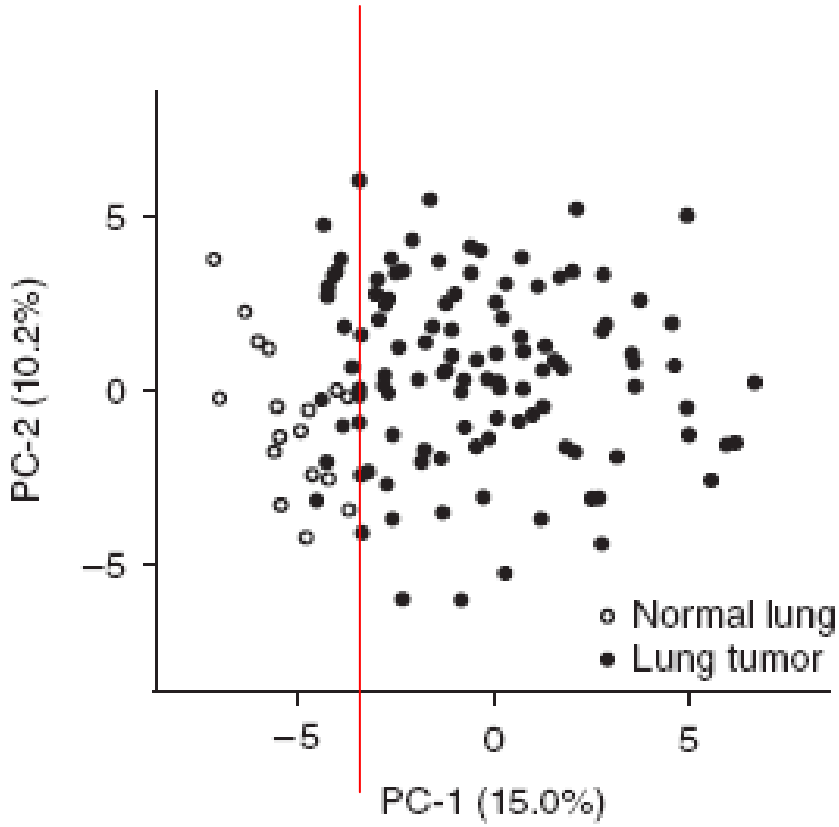
Analyses of Potti dataset

- These 198 samples represent resected malignant lung tissue from 198 individual patients with different stages of non-small cell lung cancer.

- 80 biomarker classified 90% (178 of 198) of samples.

NEJM. 355, 570–580, 2006

Fig. 3. Principal component analysis (PCA) of airway biomarker gene expression in lung tissue samples.



P= 0.026

80 biomarker probesets
Bhattacharjee dataset (n=128)

Supplementary Fig. 4. Comparison of bronchoscopy cytopathology and biomarker prediction accuracies in the primary dataset by (**a**) cancer stage or (**b**) cancer subtype.

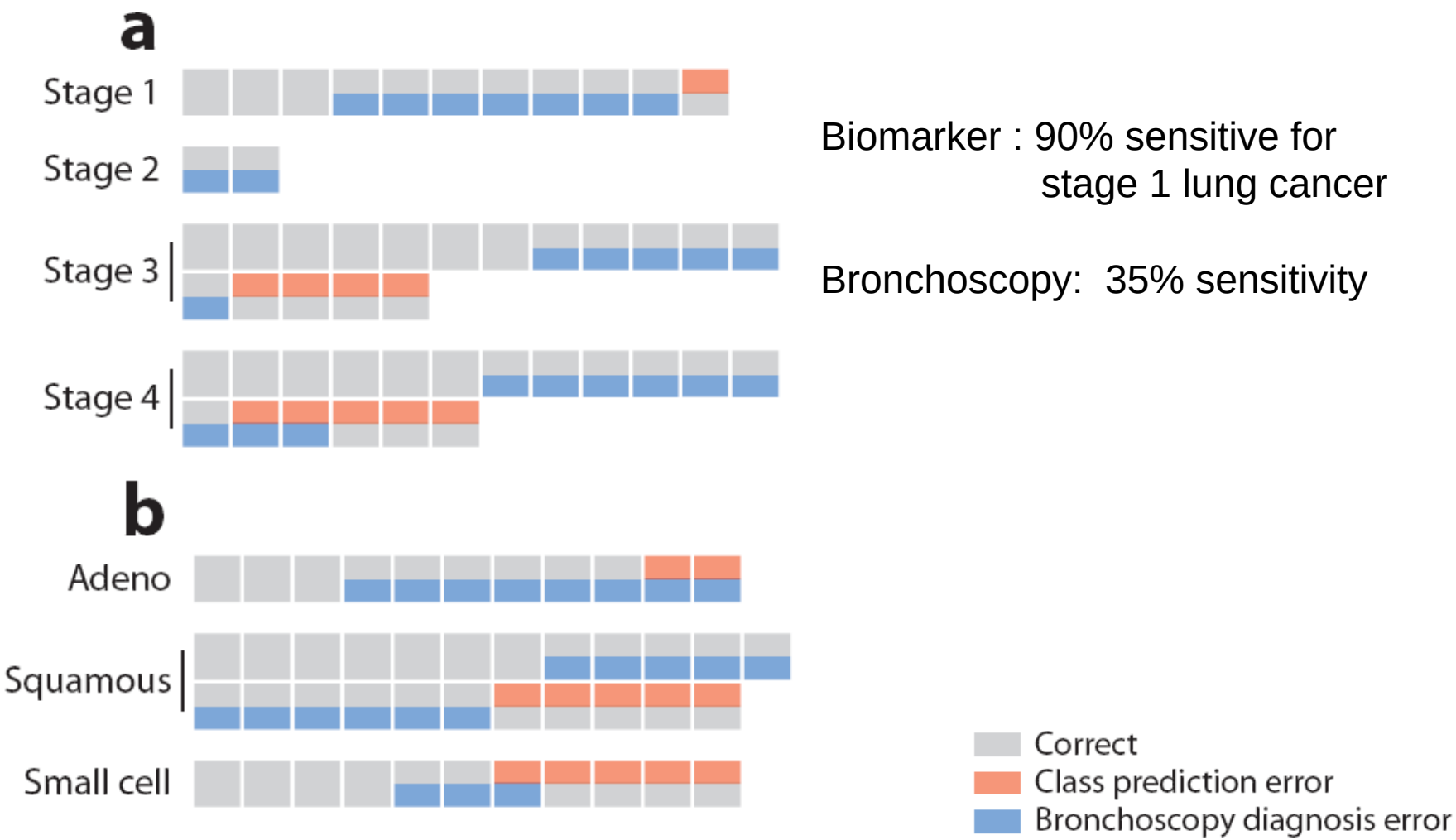


Fig. 4. Diagnostic utility of bronchoscopy and the gene-expression biomarker.

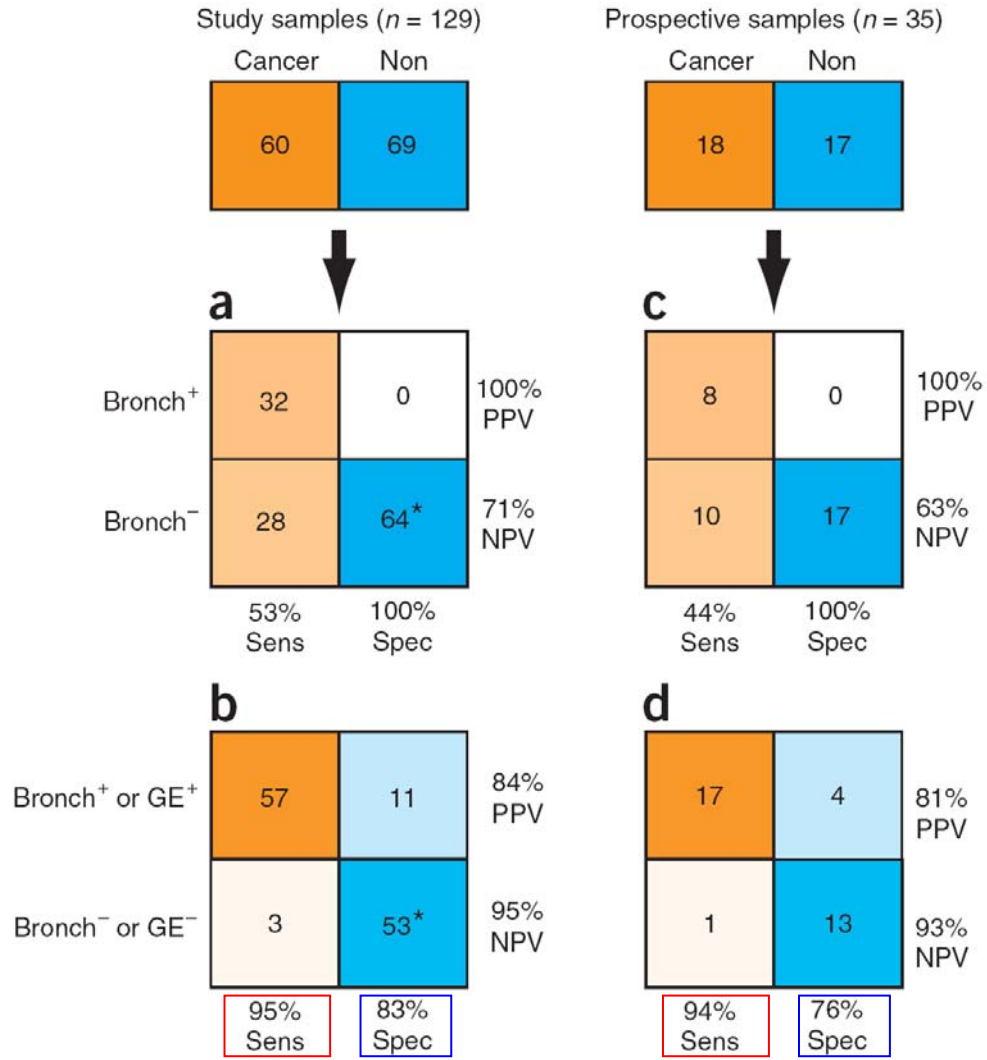
a	b
c	d

$$\text{Sensitivity} = \frac{a}{a+c}$$

$$\text{Specificity} = \frac{d}{b+d}$$

$$\text{PPV} = \frac{a}{a+b}$$

$$\text{NPV} = \frac{d}{c+d}$$



Conclusion

- Identification of airway gene-expression biomarker increased the diagnostic rate of smokers with suspect lung cancer.
- Combining cytopathology with the gene-expression biomarker improves the diagnostic sensitivity and NPV.
- These biomarkers may have the potential to identify high-risk smokers.

Thanks for your attention
and
welcome to feedback.

Expression measures

MAS 5.0•GeneChip®MAS 5.0 software uses Signal with MM*a new version of MM that is never larger than PM.•If MM < PM, MM* = MM.•If MM >= PM,–SB = TukeyBiweight(log(PM)-log(MM)) (log-ratio).–log(MM*) = log(PM)-log(max(SB, +ve)).•TukeyBiweight: B(x) = (1 – (x/c)^2)^2 if |x|<c, 0 ow.)}{log(BiweightTukey *jjMMPMsignal−=

Expression measures
RMA

Irizarry et al. (2003).

1. Estimate background BG and use only background-corrected PM: log2(PM-BG).
2. Probe level normalization of log2(PM-BG) for suitable set of chips.
3. Robust Multi-array Average, RMA, of log2(PM-BG).