Unlocking biomarker discovery: Large scale application of aptamer proteomic technology for early detection of lung cancer

Rachel M. Ostroff[1], William L. Bigbee[2], Wilbur Franklin[3], Larry Gold[1,4], Mike Mehan[1], York E. Miller[3,5], Harvey I. Pass[6], William N. Rom[7], Jill M. Siegfried[8], Alex Stewart[1], Jeffrey J. Walker[1], Joel L. Weissfeld[9], Stephen Williams[1], Dom Zichi[1], Edward N. Brody[1]

[1]*SomaLogic, 2945 Wilderness Place, Boulder, CO 80301, USA*

[2]*Department of Pathology, University of Pittsburgh School of Medicine; University of Pittsburgh Cancer Institute, Pittsburgh, PA 15232, USA*

[3]*University of Colorado Cancer Center, University of Colorado at Denver, Anschutz Medical Campus, Aurora, CO 80045, USA*

[4]*Department of Molecular, Cellular, and Developmental Biology, University of Colorado, Boulder, CO 80309, USA*

[5]*Denver Veterans Affairs Medical Center, Denver, CO 80220, USA*

[6]*Langone Medical Center and Cancer Center, New York University School of Medicine, New York, NY 10016, USA*

[7]*Sol & Judith Bergstein Professor of Medicine and Environmental Medicine Director, Division of Pulmonary, and Critical Care, and Sleep Medicine, New York University School of Medicine, 550 First Avenue New York, NY 10016, USA*

[8]*Department of Pharmacology and Chemical Biology, University of Pittsburgh School of Medicine; University of Pittsburgh Cancer Institute, Pittsburgh, PA 15232, USA*

[9]*Department of Epidemiology, University of Pittsburgh Graduate School of Public Health; University of Pittsburgh Cancer Institute, Pittsburgh, PA 15232, USA*

Lung cancer is the leading cause of cancer deaths, because ~84% of cases are diagnosed at an advanced stage[1-3]. Worldwide in 2008, ~1.5 million people were diagnosed and ~1.3 million died[4] – a survival rate unchanged since 1960. However, patients diagnosed at an early stage and have surgery experience an 86% overall 5-year survival[2,3]. New diagnostics are therefore needed to identify lung cancer at this stage. Here we present the first large scale clinical use of aptamers to discover blood protein biomarkers in disease with our breakthrough proteomic technology[5]. This multi-center case-control study was conducted in archived samples from 1,326 subjects from four independent studies of non-small cell lung cancer (NSCLC) in long-term tobacco-exposed populations. We measured >800 proteins in 15uL of serum, identified 44 candidate biomarkers, and developed a 12-protein panel that distinguished NSCLC from controls with 91% sensitivity and 84% specificity in a training set and 89% sensitivity and 83% specificity in a blinded, independent verification set. Performance was similar for early and late stage NSCLC. This is a significant advance in proteomics in an area of high clinical need.

Over the past decade the clinical utility of low-dose CT has been evaluated[6-9] with the hope that high-resolution imaging can help detect lung cancer earlier and improve patient outcomes, much as screening has done for breast and colorectal cancers[10]. Definitive conclusions about CT screening and lung cancer mortality await results from randomized trials in the US[9] and Europe[11-14]. CT can detect small, early-stage lung tumors, but distinguishing rare cancers from common benign conditions is difficult and has led to unnecessary procedures, radiation exposure, anxiety, and cost[7,15-17]. We (J.M.S., J.L.W., and colleagues) recently reported such conclusions for the Pittsburgh Lung Screening Study (PLuSS), the largest single-institution CT screening study reported to date[6].

Other types of biomarkers have also been sought[18]. Proteins are attractive because they are an immediate measure of phenotype, in contrast to DNA which provides genotype, largely a measure of disease risk[5]. However, most efforts fail to identify clinically useful biomarkers[19] because proteomic technologies have not achieved adequate coverage, sensitivity, specificity, throughput, and economy to identify biomarkers whose signals rise above the noise of sample variability and patient

comorbidities. Our new proteomic technology[5] achieves these goals and represents a platform with potentially broad application.

We designed and executed this study to current rigorous clinical biomarker study standards[20-22] with the following goals: (1) maximize biomarker robustness, validity, and reliability at the discovery phase; (2) minimize potential sample bias; (3) validate results independently. The clinical question was: "does this at-risk, tobacco-exposed individual have lung cancer?" The study was a case-control design that followed the Prospective-specimen-collection-Retrospective-Blinded-Evaluation (PRoBE) design criteria[20,22] endorsed by the U.S. National Cancer Institute's Early Detection Research Network (EDRN).

Critical study design features include: (1) clinical question and study design defined prospectively, prior to obtaining samples; (2) samples acquired from four independent study sites to minimize bias; (3) specimens collected following EDRN protocols[21] from subjects prior to diagnosis from a cohort that represents the target population for the clinical question; (4) an independent verification set as defined by current recommendations[20]; and (5) pre-defined statistical analysis plan and minimally acceptable performance criteria for sensitivity and specificity pre-defined per PRoBE design criteria[20].

The study analyzed 1,326 serum samples from four independent biorepositories: New York University (NYU)[23]; Roswell Park Cancer Institute (RPCI)[24]; The University of Pittsburgh (PITT)[6]; and a commercial biorepository (BioServe (BS)) (Supplementary Information (SI) and Supplementary Table 1). The study included patients diagnosed with pathologic or clinical stage I-III NSCLC and a high-risk control population with a history of long-term tobacco use, including active and ex-smokers with ≥10 pack-years of cigarette smoking. The control populations were selected randomly within each study to represent the patient population at risk for lung cancer that would be candidates for CT screening, with a ratio of case:control of 1:3.5.

Samples were randomly distributed into independent sets for classifier training and blinded verification (Fig. 1). Study demographics (Table 1) show no significant difference in these two sets. More than 45% of NSCLC cases were pathologically confirmed stage IA or IB or clinical stage I with adenocarcinoma representing the major

histological diagnosis (Table 2). All lung cancer patients had a biopsy-proven cancer diagnosis.

We measured the quantity of 813 proteins in each of the 1,326 samples with our proteomic platform[5]. We followed a pre-defined two-phase analysis plan to identify biomarkers and develop a classifier to distinguish lung cancer subjects from controls within the training set (training phase) and to test the classifier performance with the blinded independent verification set (verification phase). The training phase entailed two steps – biomarker selection and algorithm training with cross-validation.

To select biomarkers we performed a systematic analysis that narrowed the potential biomarker field for algorithm training to increase the probability of true discovery, yet still cast a relatively broad net. We used a Naïve Bayes (NB) method to systematically assess potential biomarker performance by preset criteria and we applied the method to subsets of the training data to broaden our cast for potential biomarkers (details in SI). The results identified a set of 44 potential biomarkers (Supplementary Table 2) that distinguish lung cancer from controls across a range of comparisons in the training set while minimizing potential "preanalytical variability" – artifacts introduced by variations in sample collection and storage[25,26].

Preanalytical variability underlies common failures to translate candidate biomarkers into clinically useful tests[19,26]. We assessed this in the study by measuring differences in protein levels within the same disease class (NSCLC or control) between different sites and comparing them to differences observed between NSCLC and control populations. The results (Fig. 2) show significant preanalytical variability between sites. However, proteins most affected by preanalytical variability are distinct from potential NSCLC biomarkers. Many proteins that exhibit preanalytical variability (Supplementary Table 4) are previously known to be susceptible to variations in sample collection and handling[25,26]. This result confirms that pre-analytical variability exists in our study and shows that our study design largely overcomes this variability to maximize the chances of discovering true, robust biomarkers of NSCLC.

To develop a potential diagnostic to distinguish NSCLC from controls, we trained NB classifiers starting with the 44 potential biomarkers we identified using a "greedy" forward search algorithm and ten-fold stratified cross validation, starting with

combinations of two biomarkers and increasing in steps of one. We constructed many high-performing eight to twelve-biomarker classifiers from this set of 44 potential biomarkers. This suggests that there is significant redundancy in the information contained within the set of potential biomarkers. Cross validated classifier performance reached a plateau with twelve biomarkers, indicating the optimal number of biomarkers for subsequent analyses. From the thousands of resulting 12-biomarker classifiers, we selected one based on pre-defined performance criteria (Supplementary Table 3) for discrimination of NSCLC from controls, sensitive detection of Stage I disease, and maintaining performance in chronic obstructive airways disease (COPD). With the training set, the classifier achieved 91% sensitivity, 84% specificity, and an area under the curve (AUC) of 0.91 (Fig. 3). The results (Table 3) show that sensitivity is maintained for Stage I NSCLC (90% for training set). The classifier performed well on samples from all four study sites (Supplementary Fig. 1).

The 12 biomarkers are shown in Table 4. The estimated serum concentrations for these markers span 4 logs (10pM-100nM). About half the control group had benign pulmonary nodules detected by CT (Table 1), and the performance of the classifier was found to be similar in that subgroup to the whole (Table 3). We also tested the effect of other attributes that could affect classifier performance such as age, smoking history, and COPD, but found little effect (Supplementary Tables 5 and 6). Age has a moderate effect on the shape of the ROC curve because the probability of cancer increases with age (Supplementary Fig. 2) but this effect can be controlled by adjusting the prior probability of cancer in the Bayes classifier model.

The classification performance of the fixed algorithm was tested on the blinded independent verification set and verified by a third party reader to achieve 89% sensitivity and 83% specificity, nearly matching the training set performance.

The biomarkers identified in this study encompass functions of cell movement, inflammation, and immune monitoring that may contribute to cancer development (SI). Some of these proteins, such as CD30 ligand, endostatin, HSP90, MIP-4, pleiotrophin, PRKCI and YES were up-regulated in lung cancer, consistent with their proposed biological roles in proliferation, invasion, or host inflammatory and immune response to the tumor (SI). We observed decreased levels of some proteins in the serum of lung

cancer patients compared to controls, including cadherin-1, LRIG3, sL-selectin, SCRsR, ERBB1 and RGM-C. Lower circulating levels of many of these proteins are associated with relief of inhibition of growth and invasion (SI).

Some of the biomarkers described in this study are the soluble domains of membrane receptors, and the function of the circulating form of these proteins may oppose their membrane-bound counterparts. For example, ERBB1 is often over-expressed in the membrane of NSCLC cells; yet, we and others have found decreased levels of the soluble domain of this protein in patient serum[27].

This study is the first large-scale application of our high-throughput, sensitive, highly multiplexed proteomic discovery platform [reported in an accompanying paper[5]] to discover and verify a novel biomarker panel for early detection of disease. The breadth of this study and the dynamic range of the proteome interrogated by our proteomic platform exceed that of other broad serum profiling platforms, including autoantibody arrays[28-30]. The biomarkers that we discovered have several potential clinical applications.

The first application is early detection of lung cancer in long-term smokers when it may be cured by surgery. Our results are a significant improvement on the performance of other recently published lung cancer biomarker studies aimed at early diagnosis[18] using mass spectrometry[23,31,32] or gene expression[33]. This performance could allow for testing of individuals with increased lung cancer risk, with subsequent CT screening based on the blood test result.

A second potential application is a test for diagnosing lung cancer in subjects with suspicious lung nodules identified by CT, which could help mitigate the problem of morbidity and cost associated with surgical interventions. CT screening reveals suspicious nodules in ~40% of long-term smokers[6,34,35], but ~97% are likely benign[6,35,36]. Protocols for managing these patients balance the risk of "watchful waiting" with definitive and costly invasive procedures. Watchful waiting monitors nodule growth by periodic follow-up CTs, but may miss the opportunity for early cures by removal of emergent small malignancies. Invasive procedures incur the risk of complications and death that arise from biopsy or futile surgical intervention for the predominant benign

lesions. This risk might be reduced by a new strategy to assess nodule volume doubling time by CT[14]. However, CT radiation itself increases cancer risk[37].

Based on the discoveries reported here, we have initiated clinical validation studies of populations at risk for lung cancer. Our goal is to develop a clinical blood test to enable an earlier diagnosis. This study is the first to be published in a sequence of successful biomarker discovery studies that we have already completed in different cancers and demonstrates the power of our proteomic technology to discover robust biomarkers in important diseases. This general approach can also be applied to discover biomarkers for many more conditions including infectious, inherited, neurological and metabolic diseases.

METHODS SUMMARY

We formally designed the study of early detection of NSCLC, including clinical question, target populations, and statistical power for sample numbers, prior to knowing what samples were available from biorepositories. Next we identified biorepositories and obtained pre-specified numbers of case and control serum samples randomly selected from those biorepository specimens that met our selection criteria. Serum samples were collected following NCI-EDRN clinical protocols. The resulting 1,326 serum samples were divided randomly into sets for training (75% of samples) and testing (25% of samples). We analyzed the samples with our new aptamer-based proteomic platform to measure the quantity of 813 proteins in each sample. We compared the resulting measurements for the training set to select potential NSCLC biomarkers, which we used to train Naïve Bayes algorithms to differentiate NSCLC from smoker controls. We selected an algorithm based on pre-defined criteria and tested its performance with the blinded test set. Test results were un-blinded and scored by a third party reader.

Full Methods and any associated references are available in the Supplementary Information.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgments

We thank our SomaLogic colleagues, in particular the assay group (Chris Bock, Evaldas Katilius, Tracy Keeney, Stephan Kraemer, Bridget Lollo, Suzanne Stratford, John Vaught, Alexey Wolfson), for making this work possible and providing valuable input to this manuscript. Lung cancer patient and PLuSS control subject accrual and annotation together with blood sample collection, processing, and storage at the University of Pittsburgh Cancer Institute was supported by NCI SPORE in Lung Cancer P50 CA090440 to JMS. Lung cancer studies at NYU were supported by grants from NCI/EDRN and from the Stephen E. Banner Fund for Lung Cancer to HIP, and a Biomarker grant from the NCI (5U01CA086137) to WR. Lung cancer studies at the Roswell Park Cancer Institute were supported in part by an NCI Cancer Center Support Grant (5P30CA016056). University of Colorado contributions to this study were supported by SPORE grant P50-CA58187 and EDRN grant U01-CA85070.

Author Contributions

R.M.O., L.G., S.W., D.Z., and E.B. designed the study. W.L.B., W.F., H.I.P., W.N.R., J.M.S., and J.L.W. provided clinical samples and interpreted results. M.M., A.S., and D.Z. analyzed the data. R.M.O. and J.J.W. wrote the manuscript with input from all authors. All authors evaluated and interpreted the analyzed data, and critically reviewed the manuscript.

Author Information

Correspondence and requests for materials should be addressed to R.M.O. (rostroff@somalogic.com) or J.J.W (jwalker@somalogic.com).

Tables

Table 1. Clinical characteristics of NSCLC cases and control training and verification sets

	Training Set (n=985)			Verification Set (n=341)		
	Cases	**Controls**	**p-value***	**Cases**	**Controls**	**p-value***
Individuals, no. (%)	213 (21.6)	772 (78.4)		78 (22.9)	263 (77.1)	
Sex, (%)						
Male	51.2	47.4		43.6	47.9	
Female	48.8	52.6	0.3305	56.4	52.1	0.5015
Age (yr)						
Mean (SD)	67.6 (9.8)	59.0 (10.2)	<0.0001	68.3 (10.2)	58.8 (9.6)	<0.0001
Control Nodule Status, no. (%)						
Benign nodule	n/a	420 (54.4)		n/a	145 (55.1)	
No nodule		222 (28.8)			75 (28.5)	
Unknown		130 (16.8)			43 (16.4)	
Smoking Status (no.)						
Current	54	421		25	150	
Ex	85	310	<0.0001	31	108	<0.0001
Never	11	6		7	3	
Unknown	63	35		15	2	
Smoking (PKY)**						
Mean (SD)	47.1 (33.7)	42.3 (24.2)	0.0258	40.9 (30.8)	42.3 (24.6)	0.7003

*For continuous data the differences were tested using t-tests. For categorical data significant differences were tested using the Pearson Chi-Squared Test for independence.

**Pack-years: product of the self reported number of packs of cigarettes smoked per day and the number of years of smoking.

Table 2. Clinical characteristics of NSCLC cases in the training and verification sets

	Training Cases (n=213)	**Verification Cases (n=78)**
Stage NSCLC*, no. (%)		
I	99 (46.5)	38 (49)
II	32 (15.0)	11 (14)
III	82 (38.5)	27 (35)
Not reported	-	2 (2)
Histology, no (%)		
Adenocarcinoma	120 (56.3)	49 (62.8)
Squamous	71 (33.3)	18 (23.1)
Large	2 (1.0)	2 (2.6)
NSCLC	20 (9.4)	9 (11.5)

*Clinical staging for 17 Stage I, 5 Stage II and 29 Stage III cases.

Table 3. Performance of Bayesian Classifier to distinguish NSCLC cases from controls

	Sensitivity (%), (95% CI)	**Specificity (%), (95% CI)**
NSCLC Cases		
Training Stage I-III	91 (87-95)	
Training Stage I	90 (84-96)	
10-fold Cross Validation	91 (87-95)	
Verification Stage I-III	89 (81-96)	
Verification Stage I	87 (78-96)	
Controls		
Training All Controls		84 (81-86)
Training Benign Nodules		82 (78-85)
10-fold Cross Validation		83 (80-86)
Verification All Controls		83 (79-88)
Verification Benign Nodules		85 (79-91)

Table 4. Twelve biomarker classifier proteins

Biomarker	**UniProt ID**	**Direction***	**Description**
Cadherin-1	P12830	down	cell adhesion, transcription regulation
CD30 Ligand	P32971	up	cytokine
Endostatin	P39060	up	inhibition of angiogenesis
HSP 90α	P07900	up	chaperone
LRIG3	Q6UXM1	down	protein binding, tumor suppressor
MIP-4	P55774	up	monokine
Pleiotrophin	P21246	up	growth factor
PRKCI	P41743	up	serine/threonine protein kinase, oncogene
RGM-C	Q6ZVN8	down	iron metabolism
SCF sR	P10721	down	decoy receptor
sL-Selectin	P14151	down	cell adhesion
YES	P07947	up	tyrosine kinase, oncogene

*Up or down regulation in NSCLC cases relative to controls.

Figures

Figure 1

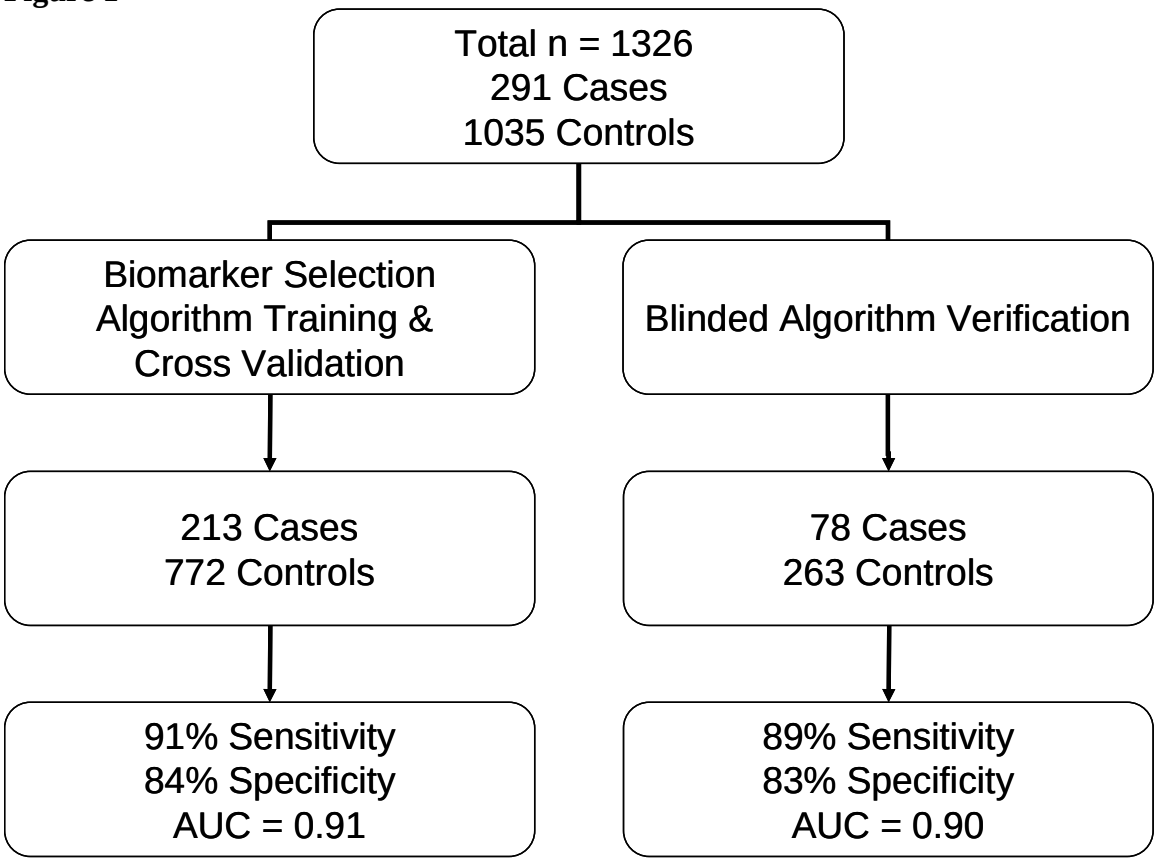

Figure 2

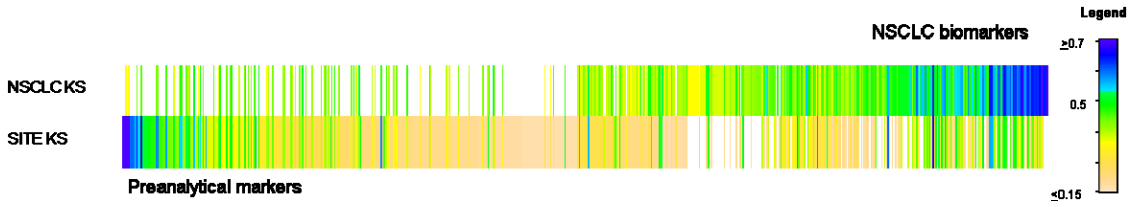

Figure 3

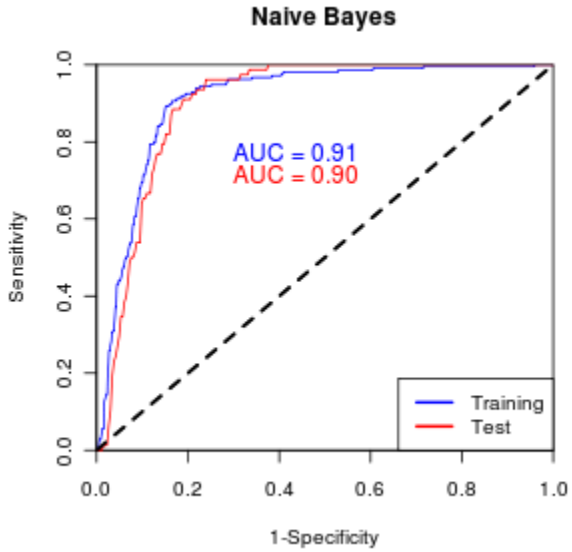

Figure Legends

Figure 1. Study flow for algorithm training and verification.

Figure 2. Heat map of the differences for protein measurements (columns) in lung cancer versus control or study site comparisons. Top row: KS distances for NSCLC versus control distributions. Bottom row: mean KS distances for all 12 pair-wise comparisons between the four sites, case and control samples analyzed separately. Proteins were ordered by subtracting the NSCLC KS distance from the mean site KS distance. This revealed groups of NSCLC biomarkers (top right) contrasting with preanalytical markers (bottom left).

Figure 3. ROC curve for 12-biomarker naïve Bayes classifier shows the true positive rate (sensitivity) and false positive rate (1-specificity) for distinguishing NSCLC cases from at-risk tobacco-exposed controls for the training set (blue) and independent verification (test) set (red).

References

1. Jemal, A. et al. Cancer statistics, 2009. *CA. Cancer J. Clin.* **59**, 225-49 (2009).
2. Okada, M. et al. Effect of tumor size on prognosis in patients with non-small cell lung cancer: the role of segmentectomy as a type of lesser resection. *J. Thorac. Cardiovasc. Surg.* **129**, 87-93 (2005).
3. Kassis, E. S. et al. Application of the revised lung cancer staging system (IASLC Staging Project) to a cancer center population. *J. Thorac. Cardiovasc. Surg.* **138**, 412-418 e1-2 (2009).
4. *World Cancer Report* (eds. Boyle, P. & Levin, B.) (International Agency for Research on Cancer (IARC), Lyon, 2008).
5. Gold, L. et al. Aptamer-based multiplexed proteomic technology for biomarker discovery. **Submitted** (2010).
6. Wilson, D. O. et al. The Pittsburgh Lung Screening Study (PLuSS): outcomes within 3 years of a first computed tomography scan. *Am. J. Respir. Crit. Care Med.* **178**, 956-61 (2008).
7. Black, W. C. Computed tomography screening for lung cancer: review of screening principles and update on current status. *Cancer* **110**, 2370-84 (2007).
8. Yau, G., Lock, M. & Rodrigues, G. Systematic review of baseline low-dose CT lung cancer screening. *Lung Cancer* **58**, 161-70 (2007).
9. NLST. (2009).
10. Smith, R. A., Cokkinides, V. & Eyre, H. J. Cancer screening in the United States, 2007: A review of current guidelines, practices, and prospects. *CA Cancer J. Clin.* **57**, 90-104 (2007).
11. Blanchon, T. et al. Baseline results of the Depiscan study: a French randomized pilot trial of lung cancer screening comparing low dose CT scan (LDCT) and chest X-ray (CXR). *Lung Cancer* **58**, 50-8 (2007).
12. Infante, M. et al. Lung cancer screening with spiral CT: baseline results of the randomized DANTE trial. *Lung Cancer* **59**, 355-63 (2008).
13. van Iersel, C. A. et al. Risk-based selection from the general population in a screening trial: selection criteria, recruitment and power for the Dutch-Belgian randomised lung cancer multi-slice CT screening trial (NELSON). *Int. J. Cancer* **120**, 868-74 (2007).
14. van Klaveren, R. J. et al. Management of lung nodules detected by volume CT scanning. *N. Engl. J. Med.* **361**, 2221-9 (2009).
15. Pinsky, P. F. et al. Diagnostic procedures after a positive spiral computed tomography lung carcinoma screen. *Cancer* **103**, 157-63 (2005).
16. Welch, H. G. et al. Overstating the evidence for lung cancer screening: the International Early Lung Cancer Action Program (I-ELCAP) study. *Arch. Intern. Med.* **167**, 2289-95 (2007).
17. Brenner, D. J. Radiation risks potentially associated with low-dose CT screening of adult smokers for lung cancer. *Radiology* **231**, 440-5 (2004).
18. Brower, V. Biomarker studies abound for early detection of lung cancer. *J. Natl. Cancer Inst.* **101**, 11-3 (2009).
19. Rifai, N., Gillette, M. A. & Carr, S. A. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat. Biotechnol.* **24**, 971-83 (2006).

20. Pepe, M. S., Feng, Z., Janes, H., Bossuyt, P. M. & Potter, J. D. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. *J. Natl. Cancer Inst.* **100**, 1432-8 (2008).
21. Tuck, M. K. et al. Standard operating procedures for serum and plasma collection: early detection research network consensus statement standard operating procedure integration working group. *J. Proteome Res.* **8**, 113-7 (2009).
22. Ransohoff, D. F. & Gourlay, M. L. Sources of bias in specimens for research about molecular markers for cancer. *J. Clin. Oncol.* **28**, 698-704 (2010).
23. Greenberg, A. K. et al. S-adenosylmethionine as a biomarker for the early detection of lung cancer. *Chest* **132**, 1247-52 (2007).
24. Ambrosone, C. B., Nesline, M. K. & Davis, W. Establishing a cancer center data bank and biorepository for multidisciplinary research. *Cancer Epidemiol. Biomarkers Prev.* **15**, 1575-7 (2006).
25. Ostroff, R. et al. The stability of the circulating human proteome to variations in sample collection and handling procedures measured with an aptamer-based proteomics array. *J. Proteomics* **73**, 649-66 (2009).
26. Zhang, Z. & Chan, D. W. Cancer proteomics: in pursuit of "true" biomarker discovery. *Cancer Epidemiol. Biomarkers Prev.* **14**, 2283-6 (2005).
27. Lemos-Gonzalez, Y., Rodriguez-Berrocal, F. J., Cordero, O. J., Gomez, C. & Paez de la Cadena, M. Alteration of the serum levels of the epidermal growth factor receptor and its ligands in patients with non-small cell lung cancer and head and neck carcinoma. *Br. J. Cancer* **96**, 1569-78 (2007).
28. Chen, G. et al. Autoantibody profiles reveal ubiquilin 1 as a humoral immune response target in lung adenocarcinoma. *Cancer Res.* **67**, 3461-7 (2007).
29. Zhong, L. et al. Using protein microarray as a diagnostic assay for non-small cell lung cancer. *Am. J. Respir. Crit. Care Med.* **172**, 1308-14 (2005).
30. Gao, W. M. et al. Distinctive serum protein profiles involving abundant proteins in lung cancer patients based upon antibody microarray analysis. *BMC Cancer* **5**, 110 (2005).
31. Yildiz, P. B. et al. Diagnostic accuracy of MALDI mass spectrometric analysis of unfractionated serum in lung cancer. *J. Thorac. Oncol.* **2**, 893-901 (2007).
32. Patz, E. F., Jr. et al. Panel of serum biomarkers for the diagnosis of lung cancer. *J. Clin. Oncol.* **25**, 5578-83 (2007).
33. Spira, A. et al. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nat. Med.* **13**, 361-6 (2007).
34. Diederich, S. et al. Screening for early lung cancer with low-dose spiral computed tomography: results of annual follow-up examinations in asymptomatic smokers. *Eur. Radiol.* **14**, 691-702 (2004).
35. Swensen, S. J. et al. Screening for lung cancer with low-dose spiral computed tomography. *Am. J. Respir. Crit. Care Med.* **165**, 508-13 (2002).
36. Croswell, J. M. et al. Cumulative incidence of false-positive results in repeated, multimodal cancer screening. *Ann. Fam. Med.* **7**, 212-22 (2009).
37. Twombly, R. Federal oversight of medical radiation is on horizon as experts face off. *J. Natl. Cancer Inst.* **102**, 514-5 (2010).

Supplementary Information

Unlocking biomarker discovery: Large scale application of aptamer proteomic technology for early detection of lung cancer

1. Sample collection

All samples were collected from study participants after obtaining informed consent under institutionally approved clinical research protocols as described[1-3]. Both case and control serum samples were collected from four centers. Three of the centers (NYU, PITT and RPMC) collected serum in red top Vacutainer tubes (Becton Dickinson, Raritan, NJ) and one center (BS) collected serum in tiger top SST Vacutainer tubes (Becton Dickinson).

All samples were allowed to clot and serum was recovered by centrifugation within 2-8 hours of collection and stored at -80°C. De-identified samples were thawed once for aliquoting prior to testing with the aptamer proteomic platform. Blood samples for cases were collected from clinic patients within four weeks of the first biopsy-proven lung cancer diagnosis and prior to removal of the tumor by a surgical procedure.

All cases used in this study were confirmed to be primary lung cancer by pathology review. NSCLC staging was assigned by pathological staging for 240 subjects and clinical staging for 51 subjects. Benign nodule controls have at least one year of follow-up data and non-malignant diagnosis. Smoker controls were asymptomatic study participants with a history of tobacco use. Smoker controls from NYU and Pitt were nodule free by CT; nodule status is unknown for this control group from RP and BS. Demographic data was collected by self-report questionnaires. Additional data for cases was acquired through clinical chart review. Pulmonary function testing was assessed by spirometry for a subset of the study participants.

Table 1. Number of samples analyzed by site

Site	Cases (n=291)	Nodule Controls (n=565)	Smoker Controls (n=470)	**Total/Site**
BS	43	0	63	106
RPCI	72	66	110	248
NYU	88	238	172	498
PITT	88	261	125	474

2. Biomarker selection

We selected 44 robust biomarkers of NSCLC for further classifier development with a strategy designed to select analytes with the highest performance in classifying NSCLC cases from controls across all study sites and that were least affected by preanalytical variables.

In the first step of this analysis, we eliminated analytes that exhibited unexpected variation compared to internal controls, due to, for example, sample instability. In this process, we chose a set of analytes that performed well in six parallel naïve Bayes (NB) classifier training scenarios using two distinct subsets of the overall population: (1) NSCLC vs. controls with benign nodules identified by CT; and (2) NSCLC vs. all other smoker controls. We used these subsets to control for possible biological variability between these populations. We analyzed each sub-population in three NB training scenarios designed to control for potential preanalytical variability between study sites. Each of the three scenarios started with a unique set of potential biomarkers selected to meet one of the following criteria for a given scenario: (1) NSCLC versus controls KS $\geq$ 0.3 for all comparisons within each of the four study sites; (2) NSCLC versus control KS $\geq$ 0.3 for comparing all sites combined; (3) both criteria one and two were met.

For each scenario, we used a greedy forward search algorithm to select subsets of potential biomarkers, build NB classifiers (SI section 7), and score their performance for classifying lung cancer and controls using the training set. In the process, this meta-heuristic approach efficiently searches classifier space to identify potential biomarkers that perform best in classification.

We used a simple measure of diagnostic performance of classifiers, the numerical sum of sensitivity + specificity, and measured the frequency with which potential biomarkers were selected by the greedy algorithm for inclusion in classifier panels with sensitivity + specificity $\geq$1.7. This step produced a set of potential biomarkers for each of the six parallel analyses. We selected the final set of biomarkers as the union of these six sets. The resulting core set of 44 potential biomarkers is shown in Table 2.

Table 2. Selected potential NSCLC biomarkers*

#	Protein Name	UniProt ID	KS	q-value	NB Freq
1	BCA-1	O43927	0.34	2.51E-17	1
2	BMP-1	P13497	0.35	3.49E-18	10
3	C1s	P09871	0.29	3.92E-13	1
4	C9	P02748	0.41	1.33E-24	6
5	Cadherin-1	P12830	0.32	1.47E-15	206
6	Calpain I	P07384 P04632	0.40	8.46E-24	72
7	Catalase	P04040	0.32	1.21E-15	2
8	CD30 Ligand	P32971	0.28	1.22E-12	51
9	CDK5/p35	Q00535 Q15078	0.27	1.34E-11	31
10	CK-MB	P12277 P06732	0.33	2.51E-16	19
11	Contactin-5	O94779	0.29	1.67E-13	3
12	Endostatin	P39060	0.28	8.48E-13	33
13	ERBB1	P00533	0.46	6.32E-31	136
14	FGF-17	O60258	0.31	6.12E-15	6
15	FYN	P06241	0.13	5.19E-04	14
16	HSP 90α	P07900	0.51	7.86E-37	85
17	HSP 90β	P08238	0.39	1.50E-22	7
18	IGFBP-2	P18065	0.36	1.87E-19	54
19	IL-15 Rα	Q13261	0.29	2.62E-13	4
20	IL-17B	Q9UHF5	0.28	1.07E-12	1
21	Importin β1	Q14974	0.40	1.31E-23	30
22	Kallikrein 7	P49862	0.31	1.79E-14	43
23	LDH-H 1	P07195	0.30	8.64E-14	3
24	Legumain	Q99538	0.28	2.52E-12	1
25	LRIG3	Q6UXM1	0.34	1.13E-17	25
26	Macrophage mannose receptor	P22897	0.37	6.21E-21	21
27	MAPK13	O15264	0.34	4.66E-18	1
28	MEK1	Q02750	0.29	2.62E-13	5
29	MetAP2	P50579	0.44	3.40E-28	7
30	Midkine	P21741	0.11	1.67E-03	7
31	MIP-4	P55774	0.29	2.69E-13	43
32	MIP-5	Q16663	0.31	1.53E-14	27
33	MMP-7	P09237	0.38	1.67E-21	36
34	NACα	Q13765	0.33	7.57E-17	5
35	NAGK	Q9UJ70	0.37	1.25E-20	5
36	Pleiotrophin	P21246	0.29	5.02E-13	107
37	PRKCI	P41743	0.41	3.81E-25	97
38	Renin	P00797	0.25	1.69E-10	2
39	RGM-C	Q6ZVN8	0.27	5.43E-12	84
40	SCF sR	P10721	0.35	6.97E-19	107
41	sL-Selectin	P14151	0.29	7.88E-13	57
42	Ubiquitin+1	P62988	0.33	4.09E-17	1
43	VEGF	P15692	0.29	5.47E-13	1
44	YES	P07947	0.28	1.73E-12	47

*Measure of the relative importance of potential biomarkers selected with KS distance (KS), KS FDR-corrected q-value (q-value), frequency for naïve Bayes (NB Freq),

A NB greedy algorithm containing 12 of these biomarkers was chosen for application to the blinded verification set based on the following, pre-defined performance criteria for algorithm training and cross-validation (Table 2).

Table 3. Criteria for algorithm performance on training and cross-validation

Criteria	Minimum Performance
Sensitivity (Stage I-III) + Specificity	1.7
Stage I Sensitivity	0.85
Cross-validation Sensitivity (Stage I-III)+ Specificity	1.7
Cross-validation Stage I Sensitivity	0.85
Severe COPD Specificity	0.65
Biomarker frequency in greedy algorithm	10

3. Preanalytical variability

Table 4. Top 20 Proteins with Preanalytical Variability

Protein	UniProt ID	Avg. KS Distance[1]
C3	P01024	0.70
C3a	P01024	0.71
C3adesArg	P01024	0.64
iC3b	P01024	0.66
C3b	P01024	0.59
LTA-4 hydrolase	P09960	0.60
EPHA3	P29320	0.47
Apo B	P04114	0.46
TrkA	P04629	0.44
HIPK3	Q9H422	0.41
Angiopoietin-1	Q15389	0.38
Coagulation Factor IXab	P00740	0.42
EF-1-γ	P26641	0.40
VEGF-D	O43915	0.39
TGF-β3	P10600	0.38
Coagulation Factor IX	P00740	0.38
C4	P0C0L4, P0C0L5	0.48
IGF-I	P01343, P05019	0.37
BMP-14	P43026	0.45
HTRA2	O43464	0.37

[1]Average KS distance for within-site, class-dependent comparisons of preanalytical variation as shown in Figure 2

4. Classifier performance by site

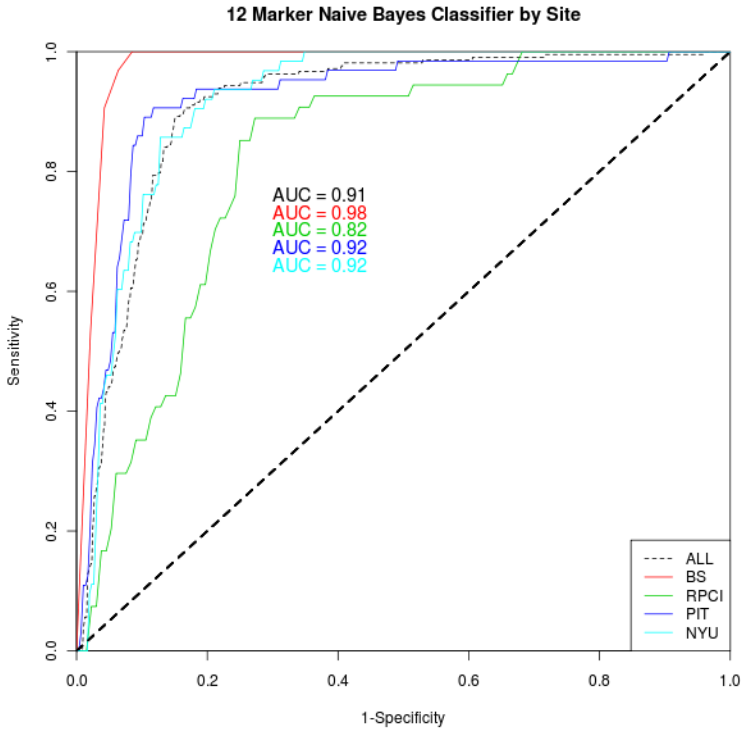

Figure 1. ROC curve performance of the 12-biomarker naïve Bayes NSCLC classifier by study site.

5. Effect of demographic attributes

To determine whether our classification results were affected either by age, smoking status, or smoking history, which are the demographics with significant differences between the case and control populations, we compared the classifier performance on subsets of the training set population divided into groups based on the median value of these attributes. The results show similar classifier performance for all subsets (Figure 2 and Table 5).

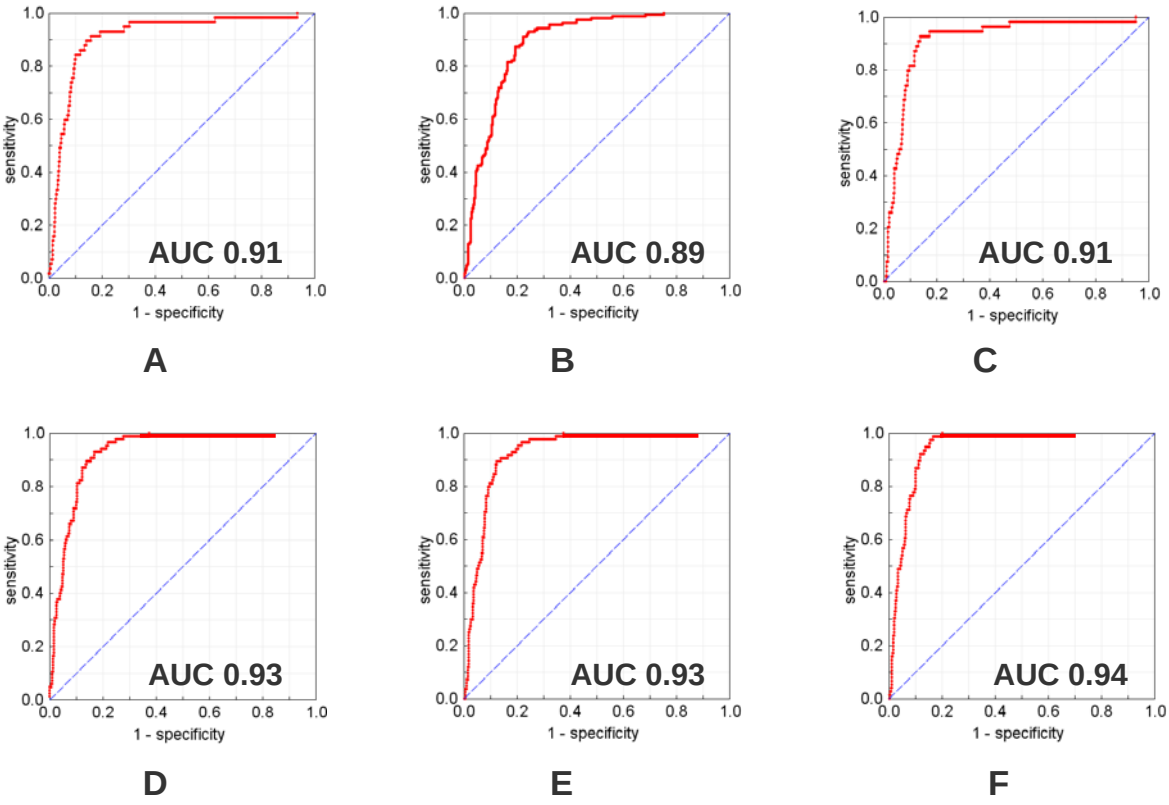

Figure 2. ROC curves show the performance of the 12-biomarker classifier for subsets of the training set population: (A) Age ≤61 (B) Age >61 (C) Current smokers (D) Ex smokers (E) ≤40 PKY (F) >40 PKY

Table 5. Performance of classifier in demographic subsets

	Cases No.	Controls No.	Sensitivity (%) (95%CI)	Specificity (%) (95%CI)	Accuracy (%) (95%CI)	AUC
Age						
≤61	57	467	84 (75-94)	89 (86-92)	88 (85-91)	0.91
>61	156	304	93 (89-97)	76 (71-80)	82 (78-85)	0.89
Smoking Status						
Current	54	421	93 (86-100)	86 (83-90)	87 (84-90)	0.91
Ex	85	310	91 (84-97)	85 (80-89)	86 (82-89)	0.93
Pack Years						
≤40	84	381	91 (84-97)	86 (83-90)	87 (84-90)	0.93
>40	76	347	97 (94-100)	84 (81-88)	87 (84-90)	0.94

To further assess whether our classification results were affected either by age, smoking status, or smoking history, we tested for potential correlation of the twelve biomarkers with these variables. The results showed no correlations except for endostatin, which showed a moderate correlation, increasing with age. This effect can be compensated for by adjusting the prior probability of cancer in the Bayes classifier model.

We also assessed the specificity of the classifier for the discrimination of controls known to have airflow obstruction (measured by GOLD score). The results are shown in Table 6. Spirometry data was incomplete for NSCLC cases, and therefore we could not calculate sensitivity.

Table 6. Classifier specificity by level of airflow obstruction

Airflow Obstruction*	FEV1 % Predicted	Number of Patients	Specificity (%), (95% CI)
GOLD 0/I	>80%	411	89 (86-92)
GOLD II	50-80%	167	84 (78-89)
GOLD III/IV	<50%	32	72 (56-87)

*Spirometric classification of airflow obstruction based on GOLD staging [4]

6. Relationship of biomarkers to tumorigenic pathways

The identified biomarkers in this study encompass functions of cell movement, inflammation, and immune monitoring that may contribute to cancer development. Some of these proteins, such as CD30 ligand, endostatin, HSP90, MIP-4, pleiotrophin, PRKCI and YES were up-regulated in lung cancer, consistent with their proposed biological roles in proliferation, invasion, or host inflammatory and immune response to the tumor. For example, CD30 ligand is a member of the TNF ligand superfamily, which stimulates T-cell growth. Up-regulation of this protein correlates with proliferation in hematological malignancies[5]. Endostatin, best known as an inhibitor of angiogenesis, has elevated serum levels in several cancers[6]. Overexpression of endostatin and its parent extracellular matrix protein, collagen XVIII have been associated with poor prognosis in NSCLC[5]. The chaperone HSP90α is important for the stability of and function of a wide range of oncoproteins, including BCR-ABL, ERBB2, EGFR, BRAF and AKT among others, and inhibitors of this protein are now in oncology clinical trials, including NSCLC[7]. HSP90

may also play a role in tumor cell resistance to complement mediated cytotoxicity[8]. MIP-4 is over-expressed in ovarian and gastric cancers, and may have a role in immunosuppression of the host tumor response[9]. Pleiotrophin is a growth factor with both mitogenic and angiogenic properties and levels in the serum of NSCLC patients have been reported to correlate with disease stage and prognosis[10]. PRKCI is an oncogene that is often amplified in NSCLC and over-expressed in lung tumors correlates with poor prognosis[11]. YES, another protein kinase and member of the src-family of tyrosine kinases, has a role in malignant transformation and increased protein levels have been reported in early stages of hepatocarcinoma[12].

We observed decreased levels of some proteins in the serum of lung cancer patients compared to controls, including cadherin-1, LRIG3, sL-selectin, SCRsR, ERBB1 and RGM-C. Lower circulating levels of many of these proteins are associated with relief of inhibition of growth and invasion. For example, cadherin-1 is critical for cell adhesion and indirectly affects transcriptional regulation circuits through β-catenin[13]. Consistent with our results, reduced expression has been reported in lung cancer, and loss of cadherin-1 is a key event leading to loss of adherence, tumorgenicity and metastasis[14]. The LRIG family consists of membrane proteins with soluble leucine rich repeat domains and immunoglobulin-like domains. Down-regulation of expression of this protein in glioblastoma cell lines resulted in increased proliferation and invasion, decreased apoptosis and increased EGFR expression, leading to the hypothesis that LRIG is a tumor suppressor[15]. L-selectin plays a role in activation of naïve lymphocytes that participate in immune surveillance and antitumor immunity. It also mediates the adherence of lymphocytes to endothelial cells. Lower expression of L-selectin may be a component of the immune suppression observed in many cancer patients[16].

Some of the biomarkers described in this study are the soluble domains of membrane receptors, and the function of the circulating form of these proteins may oppose their membrane-bound counterparts. Turner et al.[17] proposed that soluble SCF-receptors regulate kit activation. Our results suggest that a low level of SCF-sR fails to titrate SCF, which makes more SCF available for binding cancer cells. Unlike the membrane bound form, soluble RGM-C inhibits hepcidin expression[18,19]. We find that RGM-C is down

regulated in NSCLC serum, consistent with increased intracellular iron and proliferative cell growth[20].

7. Statistical Methods

Naïve Bayes. The naive Bayes classifier assumes independence between the samples, and models the distributions of the training classes to make predictions[21]. We used normal distributions to model out data, however the features in our data often contain distributions with heavy tails so maximum likelihood estimation of the distribution parameters performs poorly. We therefore modeled our distributions as log-normal distributions and used the Gauss-Newton algorithm to fit the data.

Kolmogorov-Smirnoff (KS) statistic. The KS statistic is a non-parametric measure of the difference between two distributions. The two-sample KS Statistic is $K = \sup_x |F_\alpha(x) - F_\beta(x)|$, where $F_\alpha(x)$ and $F_\beta(x)$ are empirical cumulative distributions for two populations of values.

Constructing Bayesian Classifiers. We constructed Bayesian classifiers using sets of potential biomarkers identified as described above. We used a parametric model to capture the underlying protein distribution for a given state. The simplest parametric model for the probability density function (pdf) for a single protein is a normal distribution, completely described by a mean u and variance σ^2 (Eq. 1).

$$pdf(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad (1)$$

Many protein distributions were observed to be normal with respect to the logarithm of the concentration. Figure 6 displays the numeric cdfs and their fit to a normal distribution in log concentrations x (Eq. 2).

$$cdf(x) = \int_{-\infty}^{x} pdf(y)dy \qquad (2)$$

The models fit the data well. More complex models of the probability distribution functions may be used when warranted but the simple model gives a good description of the data here.

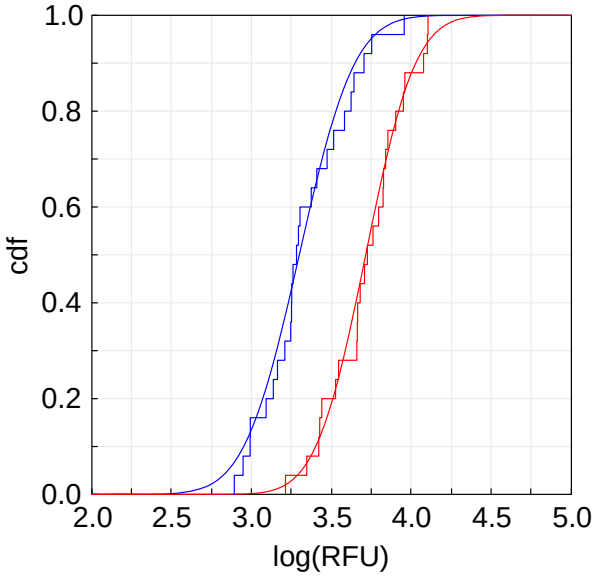

Figure 6. Curve fits of empirical cdfs of example protein (Factor H) with normal probability distribution functions for log transformed concentrations. The parameter fit was obtained through nonlinear least-squares analysis of the numeric cdfs to the normal distribution model, yielding μ of 3.3 and σ of 0.27 for the control distribution and 3.7 and 0.24 for μ and σ for the diseased distribution. The simple model fits the data extremely well.

To combine multiple markers, a multivariate normal distribution was used to model the probability density function (pdf) for each class. For n markers, the multivariate pdf is given by the following equation (Eq. 3).

$$pdf(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \qquad (3)$$

where $\mathbf{x}$ is an n-component vector of protein levels, $\boldsymbol{\mu}$ is an n-component vector of mean protein levels, $\boldsymbol{\Sigma}$ is the n x n covariance matrix and $|\boldsymbol{\Sigma}|$ and $\boldsymbol{\Sigma}^{-1}$ are its determinant and inverse. In its simplest form, we can assume a diagonal representation for $\boldsymbol{\Sigma}$. Such an approximation leads to a naïve Bayes model, which assumes independence between the markers. In this work, we exclusively use the naïve Bayes model for constructing classifiers. The parameter values for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ used in the naïve Bayes classification were obtained from nonlinear regression analysis as described above.

Greedy Algorithm for Classifier Generation. The addition of subsequent markers with good KS distances will, in general, improve the classification performance if the subsequently added markers are independent of the first marker. Using the sensitivity (fraction of true positives) plus specificity (fraction of true negatives) as a classifier score, it is straightforward to generate many high scoring classifiers with a variation of a greedy algorithm. A greedy algorithm is any algorithm that follows the problem solving meta-heuristic of making the locally optimal choice at each stage with the hope of finding the global optimum.

The algorithm approach used here is described as follows. All single analyte classifiers are generated from a table of potential biomarkers and added to a list. Next, all possible additions of a second analyte to each of the stored single analyte classifiers is then performed, saving a predetermined number of the best scoring pairs, say, for example, a thousand, on a new list. All possible three marker classifiers are explored using this new list of the best two-marker classifiers, again saving the best thousand of these. This process continues until the score either plateaus or begins to deteriorate as additional markers are added.

Supplementary References

1. Greenberg, A. K. et al. S-adenosylmethionine as a biomarker for the early detection of lung cancer. *Chest* **132**, 1247-52 (2007).
2. Ambrosone, C. B., Nesline, M. K. & Davis, W. Establishing a cancer center data bank and biorepository for multidisciplinary research. *Cancer Epidemiol. Biomarkers Prev.* **15**, 1575-7 (2006).
3. Wilson, D. O. et al. The Pittsburgh Lung Screening Study (PLuSS): outcomes within 3 years of a first computed tomography scan. *Am. J. Respir. Crit. Care Med.* **178**, 956-61 (2008).
4. GOLD, C. (2008).
5. Iizasa, T. et al. Overexpression of collagen XVIII is associated with poor outcome and elevated levels of circulating serum endostatin in non-small cell lung cancer. *Clin. Cancer Res.* **10**, 5361-6 (2004).
6. Suzuki, M. et al. Serum endostatin correlates with progression and prognosis of non-small cell lung cancer. *Lung Cancer* **35**, 29-34 (2002).
7. Banerji, U. Heat shock protein 90 as a drug target: some like it hot. *Clin. Cancer Res.* **15**, 9-14 (2009).
8. Gancz, D. & Fishelson, Z. Cancer resistance to complement-dependent cytotoxicity (CDC): Problem-oriented research and development. *Mol. Immunol.* **46**, 2794-800 (2009).

9. Schutyser, E., Richmond, A. & Van Damme, J. Involvement of CC chemokine ligand 18 (CCL18) in normal and pathological processes. *J. Leukoc. Biol.* **78**, 14-26 (2005).
10. Jager, R. et al. Serum levels of the angiogenic factor pleiotrophin in relation to disease stage in lung cancer patients. *Br. J. Cancer* **86**, 858-63 (2002).
11. Erdogan, E., Klee, E. W., Thompson, E. A. & Fields, A. P. Meta-analysis of oncogenic protein kinase Ciota signaling in lung adenocarcinoma. *Clin. Cancer Res.* **15**, 1527-33 (2009).
12. Nonomura, T. et al. Identification of c-Yes expression in the nuclei of hepatocellular carcinoma cells: involvement in the early stages of hepatocarcinogenesis. *Int. J. Oncol.* **30**, 105-11 (2007).
13. Ceteci, F. et al. Disruption of tumor cell adhesion promotes angiogenic switch and progression to micrometastasis in RAF-driven murine lung cancer. *Cancer Cell* **12**, 145-59 (2007).
14. Charalabopoulos, K., Gogali, A., Kostoula, O. K. & Constantopoulos, S. H. Cadherin superfamily of adhesion molecules in primary lung cancer. *Exp. Oncol.* **26**, 256-60 (2004).
15. Cai, M. et al. Inhibition of LRIG3 gene expression via RNA interference modulates the proliferation, cell cycle, cell apoptosis, adhesion and invasion of glioblastoma cell (GL15). *Cancer Lett.* **278**, 104-12 (2009).
16. Hanson, E. M., Clements, V. K., Sinha, P., Ilkovitch, D. & Ostrand-Rosenberg, S. Myeloid-derived suppressor cells down-regulate L-selectin expression on CD4+ and CD8+ T cells. *J. Immunol.* **183**, 937-44 (2009).
17. Turner, A. M. et al. Identification and characterization of a soluble c-kit receptor produced by human hematopoietic cell lines. *Blood* **85**, 2052-8 (1995).
18. Babitt, J. L. et al. Bone morphogenetic protein signaling by hemojuvelin regulates hepcidin expression. *Nat. Genet.* **38**, 531-9 (2006).
19. Babitt, J. L. et al. Modulation of bone morphogenetic protein signaling in vivo regulates systemic iron balance. *J. Clin. Invest.* **117**, 1933-9 (2007).
20. Ward, D. G. et al. Increased hepcidin expression in colorectal carcinogenesis. *World J. Gastroenterol* **14**, 1339-45 (2008).
21. Duda, O., Hart, P. E. & Stork, D. G. *Pattern Classification* (John Wiley and Sons, New York, 2001).