Gene Profiling of Clinical Routine Biopsies and Prediction of Survival in Non–Small Cell Lung Cancer

Florent Baty[1], Michaël Facompré[2], Sergio Kaiser[3], Martin Schumacher[3], Miklos Pless[4], Lukas Bubendorf[5], Spasenija Savic[5], Estelle Marrer[3], Wolfgang Budach[3], Martin Buess[6], Jeanne Kehren[3], Michael Tamm[7], and Martin H. Brutsche[1]

[1]Department of Pneumology, Kantonsspital St. Gallen, St. Gallen, Switzerland; [2]Department of Biomedicine, [5]Institute for Pathology, and [7]Department of Pneumology, University Hospital Basel, Basel, Switzerland; [3]Biomarker Development, Novartis AG, Basel, Switzerland; [4]Medical Oncology, Kantonsspital Winterthur, Winterthur, Switzerland; and [6]Department of Oncology, Claraspital Basel, Basel, Switzerland

***Rationale*: Global gene expression analysis provides a comprehensive molecular characterization of non–small cell lung cancer (NSCLC).**

***Objectives*: To evaluate the feasibility of integrating expression profiling into routine clinical work-up by including both surgical and minute bronchoscopic biopsies and to develop a robust prognostic gene expression signature.**

***Methods*: Tissue samples from 41 chemotherapy-naive patients with NSCLC and 15 control patients with inflammatory lung diseases were obtained during routine clinical work-up and gene expression profiles were gained using an oligonucleotide array platform (NovaChip; 34′207 transcripts). Gene expression signatures were analyzed for correlation with histological and clinical parameters and validated on independent published data sets and immunohistochemistry.**

***Measurements and Main Results*: Diagnostic signatures for adenocarcinoma and squamous cell carcinoma reached a sensitivity of 80%/80% and a specificity of 83%/94%, respectively, dependent on the proportion of tumor cells. Sixty-seven of the 100 most discriminating genes were validated with independent observations from the literature. A 13-gene metagene refined on four external data sets was built and validated on an independent data set. The metagene was a strong predictor of survival in our data set (hazard ratio = 7.7, 95% CI [2.8–21.2]) and in the independent data set (hazard ratio = 1.6, 95% CI [1.2–2.2]) and in both cases independent of the International Union against Cancer staging. Vascular endothelial growth factor-β, one of the key prognostic genes, was further validated by immunohistochemistry on 508 independent tumor samples.**

***Conclusions*: Integration of functional genomics from small bronchoscopic biopsies allows molecular tumor classification and prediction of survival in NSCLC and might become a powerful adjunct for the daily clinical practice.**

AT A GLANCE COMMENTARY

Scientific Knowledge on the Subject

Gene expression microarrays have been successfully used for the diagnosis and prognosis of patients with non–small cell lung cancer (NSCLC). However, most studies investigated tumor samples obtained from surgical specimens only, which limits their findings to operable early-stage patients.

What This Study Adds to the Field

The proposed strategy of integrating functional genomics from small bronchoscopic biopsies allows molecular tumor classification and prediction of survival in patients with NSCLC of all stages and has the potential to become a powerful adjunct for the daily clinical practice.

Very recently the distinction between different histological subtypes of non–small cell lung cancer (NSCLC) began to become relevant as different agents showed preferential activities according to the histological subtype (1). Targeted therapies for NSCLC based on molecular markers start to play a more and more prominent role (2). Early studies using high-throughput gene expression technology aimed to identify gene classifiers of lung cancer subtypes (3, 4) or predictors for disease outcome (5, 6). These studies provided an important contribution regarding the identification of distinct subtypes among adenocarcinomas (AC) (3, 5) and squamous cell carcinomas (SCC) (7, 8) associated with specific gene expression patterns that correlated with survival (4). Recent studies described gene signatures predicting survival with good accuracy after validation in independent data sets (9–12). All gene expression microarray studies in lung cancer published so far are based on tumor samples obtained during lung cancer surgery with curative intent and some include CT-guided transthoracic biopsies (13, 14). As a consequence, most of these studies focused on early-stage NSCLC (7, 9). However, the vast majority of patients have more advanced disease (15) and the findings in the studies mentioned above are not necessarily applicable to approximately 70% of the patients with lung cancer who are not surgical candidates. Spira and colleagues (16) recently evaluated the diagnostic value of functional genomics of bronchial airway epithelial cells obtained with an endoscopic cytobrush technique in smokers with suspicion of lung cancer. They could build a diagnostic signature containing 80 genes for the identification of patients with lung cancer with a sensitivity of 80% and specificity of 84%.

Flexible bronchoscopy is safe and applicable to patients with both early and advanced disease (17). Bronchoscopy represents a cornerstone of the standard clinical work-up of patients with suspected lung cancer (18). Thus, a strategy of using bronchoscopic biopsies for microarray studies together with lung biopsies derived from lung cancer resections in patients with resectable early-stage NSCLC would allow its application in almost every patient and could easily be implemented in the standard clinical work-up (19).

Whether gene expression profiling of combined small bronchoscopic biopsies taken during flexible bronchoscopy and large biopsies obtained during lung cancer surgery can help to refine diagnosis, treatment, and prognosis in patients with NSCLC is

(*Received in original form December 2, 2008; accepted in final form October 14, 2009*)

Supported by an unconditional grant from Novartis.

Correspondence and requests for reprints should be addressed to Prof. Martin H. Brutsche, M.D., Ph.D., Pneumologie, Kantonsspital St. Gallen, CH-9007 St. Gallen, Switzerland. E-mail: martin.brutsche@kssg.ch

This article has an online supplement, which is accessible from this issue's table of contents at www.atsjournals.org

Am J Respir Crit Care Med Vol 181. pp 181–188, 2010
Originally Published in Press as DOI: 10.1164/rccm.200812-1807OC on October 15, 2009
Internet address: www.atsjournals.org

unknown. Therefore, the aim of this study was to evaluate the feasibility and validity of gene expression profiling for molecular classification and prediction of survival from biopsies obtained during the standard clinical work-up and including bronchoscopic tumor biopsies. Some of the results of these studies have been previously reported in the form of an abstract (20).

METHODS

Population and Tissue Samples

Eighty-five patients underwent flexible video bronchoscopy for suspicion of lung cancer in the Department of Pneumology, University Hospital Basel, Switzerland, from November 2002 to November 2005 and were willing to give informed consent to have additional bronchial biopsies for gene expression analyses. Of these, 10 patients with early-stage disease underwent surgery, during which a tumor biopsy was taken for gene expression analysis. When available, the surgical biopsies were taken for further analyses. Fifteen patients with small cell lung cancer (SCLC) and nine patients with prior chemotherapy were excluded from the study. Five cases were excluded because of insufficient RNA quality. In total 56 patients (79 samples) consisting of 41 patients with confirmed NSCLC and 15 control patients, in whom the diagnosis turned out to be nonmalignant lung disease, were included in the study (Table 1). The control patients suffered from chronic inflammatory or interstitial lung diseases (interstitial lung disease: n = 4; chronic bronchitis/COPD: n = 7; sarcoidosis: n = 4; 9/15 current smokers). The study was approved by the local ethical review board (Ethik-Kommission beider Basel EKBB No. 214/02 and 05/06).

Bronchoscopic biopsies were obtained during flexible video bronchoscopy performed according to general clinical standards in conscious sedation with either midazolam or propofol. The premedication generally included local anesthesia of nose, mouth, and oropharynx, as well as hydrocodone (5 mg i.v.). During the whole procedure, the patients received oxygen at a flow rate of 2 L/min with a cannula or face mask. Heart rate, blood pressure, and oxygen saturation were monitored. Biopsies for the standard clinical purposes were sampled first. Then, two additional biopsies (diameter approximatively 0.5 mm each) were sampled for the current study. These were directly put into RNA*later* (Ambion, Austin, TX), and placed into a −20°C freezer within 1 hour, until transferred and stored at −80°C. Surgical specimens were directly dissected in the operating theater and typically a biopsy of 1 cm^3 was conserved for later RNA isolation.

Histopathologic Evaluation

Biopsies for gene expression analysis and histopathological evaluation were taken at the same endobronchial site during the same procedure. Biopsies for histopathologic evaluation were fixed in 4% buffered formalin, paraffin-embedded, cut at 4 μm, and stained with hematoxylin and eosin, Alcian blue periodic acid Schiff, and elastica van Gieson according to routine procedures. All biopsies were evaluated by two independent lung pathologists for the following characteristics: (*1*) tumor quantity in percent of the whole tissue section, (*2*) histological type (SCC, AC, large cell carcinoma with and without neuroendocrine differentiation, or not otherwise specified NSCLC), (*3*) histological grade, (*4*) presence and type of inflammatory cells (mononuclear cells, neutrophils, eosinophils, granulomas), and (*5*) presence and quantity of necrosis.

Gene Expression Microarray Experiments

RNA was extracted by an optimized TRIzol Reagent (Invitrogen, Carlsbad, CA) protocol followed by the RNA purification steps of the Qiagen RNeasy Microkit including on-column DNase treatment. RNA extraction samples were hybridized on highly sensitive NovaChip-microarrays (Novartis, Basel, Switzerland), an evanescent resonator platform (21). A Tecan HS 4800 Hybridization station was used for prewash/hybridization/postwash of the microarrays (for more details *see* online supplement).

Data Analysis

Data were normalized by scaling the intensity distribution using the 75% trimmed mean, and variance stabilized by logarithmic transformation. Technical batch effects were adjusted using the Partek batch remover. The supervised between-groups analysis (22, 23) was used to identify the genes that best discriminate between the disease categories. A gene classifier was built using a genetic algorithm combined with the nearest centroid classification method (implemented in the R package GALGO) (24). To assess the unbiased prediction accuracy of the classifier, the procedure includes an internal routine of two-level cross-validation, which consists of splitting the initial data set several times into two-thirds training sets and one-third test sets. Each first-level training set is thereafter divided into many second-level training/test sets.

Survival analysis was performed by applying univariate Cox proportional hazards regression and supervised principal component analysis (25). A metagene based on a linear combination of the genes that best predict survival was built according to the procedure described by Bair and Tibshirani (25). Based on the median of the metagene scores, we built a binary score (low/high risk) and displayed the survival results by using Kaplan-Meier curves. The entire microarray data set is available

TABLE 1. PATIENT AND SAMPLE DESCRIPTION

	Control	NSCLC			
	Chronic Inflammatory Lung Disease (*N = 15*)	NSCLC, Total (*N = 41*)	Adenocarcinoma (*N = 13*)	Squamous Cell Carcinoma (*N = 14*)	NSCLC, Not Otherwise Specified (*N = 14*)
Age, yr	57 (34–82)	66 (42–80)	68 (52–80)	62 (54–78)	68 (42–80)
Male sex, no. (%)	11 (73)	27 (66)	6 (46)	12 (86)	9 (64)
Stage of cancer, no. (%)					
I	—	7 (17)	3 (23)	2 (14)	2 (14)
II	—	8 (20)	1 (8)	5 (36)	2 (14)
III	—	9 (22)	1 (8)	4 (29)	4 (29)
IV	—	17 (41)	8 (62)	3 (21)	6 (43)
Follow-up, mo	—	22 (14–44)	20 (14–44)	22 (17–40)	23 (19–33)
Survival, mo	—	14 (9 to >44)	13 (9 to >44)	26 (9 to >39)	12 (5.6 to >33)
Bronchoscopic biopsy, no. of samples	15	31	7	12	12
Surgical biopsy, no. of samples	0	10	6	2	2
Tumor proportion, no. of samples (%)					
0%	—	8 (20)	4 (31)	1 (7)	3 (21)
<50%	—	7 (17)	1 (8)	3 (21)	3 (21)
>50%	—	21 (51)	7 (54)	8 (57)	6 (43)
Unknown	—	5 (12)	1 (8)	2 (14)	2 (14)

Definition of abbreviations: NSCLC = non–small cell lung cancer.
Age, follow-up, and survival are given as median (range).

online (www.ncbi.nlm.nih.gov/projects/geo/) under the data series accession number GSE11117. Four recently published independent lung cancer data sets were used to refine the construction of the metagene (3, 5, 8, 26) and another more recent independent data set (11) was used for external validation (*see* the online supplement for details).

Tissue Microarrays

The prognostic value of vascular endothelial growth factor-β (*VEGFB*) was validated on a protein level using immunohistochemistry (sc-1878; Santa Cruz Biotechnology, Inc., Santa Cruz, CA) on tissue microarrays (27) with tumor samples from 508 patients with a median follow-up of 51 (range 0–200) months (*see* online supplement).

RESULTS

Patient Population and Sample Characteristics

The characteristics of the 56 patients are reported in Table 1 (for more details *see* Table E1 in the online supplement). The NSCLC and inflammatory control groups were matched for age and sex. Seventeen percent, 20%, 22%, and 41% of patients with NSCLC were in International Union against Cancer, 6th edition (UICC) stages I, II, III, and IV, respectively. In 20% of patients with NSCLC, the lung pathologists were unable to identify tumor cells in the bronchoscopic biopsy. In these individuals the diagnosis of NSCLC was made by means of cytology of the bronchial washing, mediastinoscopy, or CT-guided biopsy. Bronchoscopic biopsies from SCC and AC provided detectable tumor cells in 92% and 33% of samples, respectively. The median survival of patients with NSCLC was 14 months (range 9 to > 44 mo) with a median follow-up of 22 months (range 14–44 mo). There were no bronchoscopy- or biopsy-related complications.

Molecular Classification

Genes discriminating control, AC, and SCC were identified with between-group analysis (Figure 1). Sixty-seven of the 100 most discriminating genes were already described in the literature as being associated with NSCLC (*see* Table E2). SCC typically exhibited an up-regulation of keratin genes; genes encoding for epithelial development, such as Ca-binding proteins; small proline-rich proteins; and antioxidant proteins, such as aldo-keto reductases. AC showed increased transcriptional levels of markers routinely used for the diagnosis of lung adenocarcinomas, such as surfactant proteins (SFTP) and napsin A aspartic peptidase (NAPSA). The overall sensitivity and specificity of the molecular classification were 0.80 and 0.89, respectively.

Restricting the analysis to bronchoscopic biopsies resulted in comparable overall prediction accuracy (sensitivity of 77% and specificity of 77%). However, because of the small number of cases obtained after restricting the analysis to bronchoscopic samples only, there was a decrease of the prediction accuracy of adenocarcinoma (sensitivity of 23%, specificity of 87%).

Six duplicated samples were available for molecular classification. Over the 100 most discriminating genes, the intraindividual variability was found to be comparatively low (median coefficient of variation [CV], 12.2%) and a good diagnostic repeatability was obtained (Figure E5A).

Prediction of Survival

The UICC stage, but not age, sex, or histology, was the only significant clinical predictor for survival ($P < 0.001$) and thus was included as a covariable in the Cox proportional hazards regression models. The development of our metagene followed these three steps: (*1*) build a metagene based on our training data set, (*2*) refine the metagene by using four external data sets, and (*3*) test the metagene in an additional independent data set. A metagene including 44 genes (34 risk genes, 10 protective genes) gave the most accurate prediction of survival ($P < 0.001$; Table 2, Figure E4). For 10 (Tomida and colleagues) (8) to 28 (Bild and colleagues) (26) of these 44 genes, corresponding transcripts were identified in at least one of the four independent data sets. Thirteen (10 risk genes, 3 protective genes) were significantly associated with survival in at least one of the four independent data sets (Table 2). These 13 genes were combined into a metagene, which provided independent prognostic information complementary to the UICC stage information ($P < 0.001$; Figures 2A, 2B, and 2E). This metagene was particularly efficient to identify patients with a survival of less than 1 year (sensitivity 78%, specificity 89%). The nomogram presented in Figure 2G describes how both variables can be used in association to predict the probability of

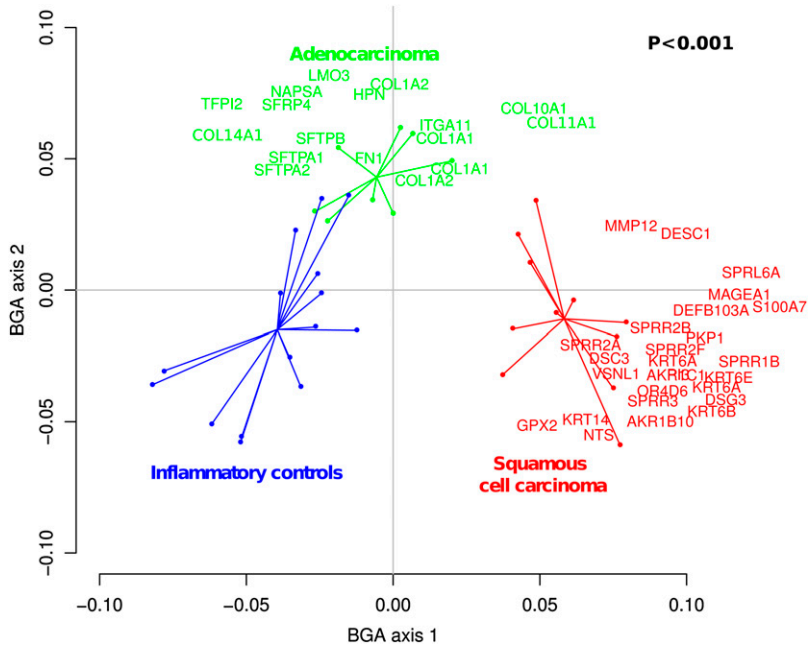

Figure 1. Molecular classification of non–small cell lung cancer based on gene expression profiles. The figure shows the classification of patients by a between-groups analysis (biplot representation). The discrimination between the three phenotypes was significant (Monte Carlo permutation test, $P < 0.001$). The most discriminating genes have the highest absolute scores on the between-group analysis axes (i.e., a large distance from the origin of the plot). Genes particularly associated with squamous cell carcinoma and adenocarcinoma are given in red and green, respectively. BGA = between-groups analysis.

TABLE 2. DESCRIPTION OF THE 44 PROGNOSTIC GENES OF THE METAGENE ASSOCIATED WITH SURVIVAL

Operon ID	HUGO Short Name	Gene Name	Chr. Position	Hazard Ratio*	*P* Value†	Bhattacharjee *et al.* (2001) (2)	Beer *et al.* (2002) (4)	Tomida *et al.* (2004) (7)	Bild *et al.* (2006) (22)
H300013783	***ARPC2***	actin-related protein 2/3 complex, subunit 2	2q36.1	6.27	6.6E-03	**0.025**	0.56	0.16	0.08
H200017863	*FOXQ1*	forkhead box q1	6p25	6.26	5.7E-04				0.26
H200012139	***SDF2***	stromal cell-derived factor 2	17q11.2	5.66	6.9E-03	0.33	**0.017**	0.11	0.88
H200009657	*MOSPD3*	motile sperm domain containing 3	7q22	0.18‡	3.4E-03				0.08
H200017969	*PMS2L9*	postmeiotic segregation increased 2-like3	7q11.23	0.19‡	4.0E-04				
H200014047	*MUT*	methylmalonyl coenzyme a mutase	6p21	5.08	4.2E-04	0.49	0.54		0.46
H200008119	*NOC4*	neighbor of cox4	16q24	4.61	4.8E-04				
H300003690	*NA*	—		4.59	3.2E-03				
H200007113	*NDUFC1*	nadh dehydrogenase (ubiquinone) 1	4q28.2	4.55	9.5E-03	0.58			0.32
H300021867	*KIAA2010*	kiaa2010	14q32	4.32	2.1E-02				0.24
H200005989	***AP3D1***	adaptor-related protein complex 3, delta 1 subunit	19p13.3	4.31	1.5E-03	**0.013**	0.53		**0.002**
H300005457	*XP_372900.2*	—	22q13.2	4.16	2.3E-03				
H200014959	***MRPL44***	mitochondrial ribosomal protein l44	2q36.1	4.13	1.8E-03				**0.02**
H300018776	*CHCHD2*	coiled-coil-helix-coiled-coil domain containing 2	7p11.2	4.10	2.1E-03				0.98
H300008134	*SLC37A2*	solute carrier family 37, member 2	11q24.2	0.24‡	2.5E-02				0.29
H200017073	*ACTR10*	actin-related protein 10 homolog (s. cerevisiae)	14q23.1	4.07	5.1E-03				0.75
H300012698	*Q6PIE2*	Kiaa0220-like protein	16p11.2	0.25‡	3.0E-03				
H200014044	*TCEB3*	transcription elongation factor b, polypeptide 3	1p36.1	3.97	1.5E-02	0.088	0.72	0.18	0.07
H200004476	*INO80*	ino80 homolog (yeast)	15q15.1	3.91	5.2E-02				
H200006915	***MYO1E***	myosin ie	15q21	3.90	9.9E-03	0.29	0.3	0.99	**0.04**
H300012071	*LOC90410*	intraflagellar transport protein ift20	17q11.2	3.87	3.1E-02				
H300013306	*APG5L*	apg5 autophagy 5-like (s. cerevisiae)	6q21	3.73	5.6E-03				
H300007836	*MGC33302*	hypothetical protein mqc33302	4q28.1	3.71	5.7E-03			0.7	0.31
H200003249	*FTSJ1*	ftsj homolog 1 (e. coli)	Xp11.23	3.70	1.2E-02	0.2			0.67
H200006553	***VEGFB***	vascular endothelial growth factor b	11q13	0.28‡	9.0E-04	**0.022**	**0.002**	**0.002**	0.054
H200014122	*RNF103*	ring finger protein 103	2p11.2	3.60	1.8E-02	0.7	0.75		0.51
H200013667	*CYB561D2*	cytochrome b-561 domain containing 2	3p21.3	3.59	6.5E-03	0.082			0.53
H200017355	***OPTN***	optineurin	10p13	0.28‡	3.4E-03	**0.016**		0.63	0.62
H300010368	*MAT2B*	dtdp-4-keto-6-deoxy-d-glucose 4-reductase	5q13.2	0.28‡	1.7E-02				
H200011731	***HEBP2***	heme binding protein 2	6q24	3.50	4.3E-03	**0.016**			0.77
H300003194	*LRRC9*	leucine rich repeat containing 9	14q23.1	3.47	8.1E-04				
H300010348	*NA*	—		3.44	9.2E-03				
H300013688	***CSNK1A1***	casein kinase 1, alpha 1	5q32	3.43	2.6E-03	0.84	0.5	0.22	**0.04**
H300015410	*NA*	—		3.41	1.3E-02				
H200004109	***CLIP1***	restin	12q24.3	3.35	3.6E-03	0.89	0.6		**0.0004**
H200009361	***MUS81***	mus81 endonuclease homolog (yeast)	11q13	0.30‡	1.5E-02				**0.0003**
H200008089	***ARG2***	arginase, type ii	14q24.1	3.32	2.6E-03	0.46	**0.005**		0.78
H200014666	***SNAP29***	synaptosomal-associated protein	22q11.21	3.29	4.7E-02			**0.048**	0.11
H300006812	*LOC91661*	hypothetical protein bc001610	19q13.42	0.30‡	2.7E-02				
H300018503	*NYREN18*	nedd8 ultimate buster-1	7q36	0.31‡	6.1E-03				
H200004564	*MGC10067*	ubiquitin-like domain containing ctd phosphatase 1	5q33.3	3.22	1.1E-02				
H300021626	*RMI1*	chromosome 9 open reading frame 76	9q21.32	3.19	2.5E-03				
H200006229	*MBTPS1*	membrane-bound transcription factor peptidase, site 1	16q24.1	3.17	1.2E-02	0.57	0.95	0.74	0.48
H300022216	*CBWD2*	hypothetical protein from clone 1659351	2q14.1	3.13	7.6E-03				0.43

Definition of abbreviations: Chr. = chromosomal; NA = not available; UICC = International Union Against Cancer.

The genes were refined using four independent data sets (the boldfaced genes were included in the 13-gene refined metagene). The *P* values presented in the last four columns were obtained by testing (log-ratio test of a univariate Cox proportional hazards model) the association between each of these genes and survival.

* Hazard ratios are reported for each gene included in the metagene. Cox proportional hazards models were fitted gene by gene including the information of UICC stages as single covariate.

† *P* values for the hazard ratios were estimated by Cox regression analysis.

‡ Protective genes have a hazard ratio between 0 and 1. Genes with a hazard ratio greater than 1 are risk genes.

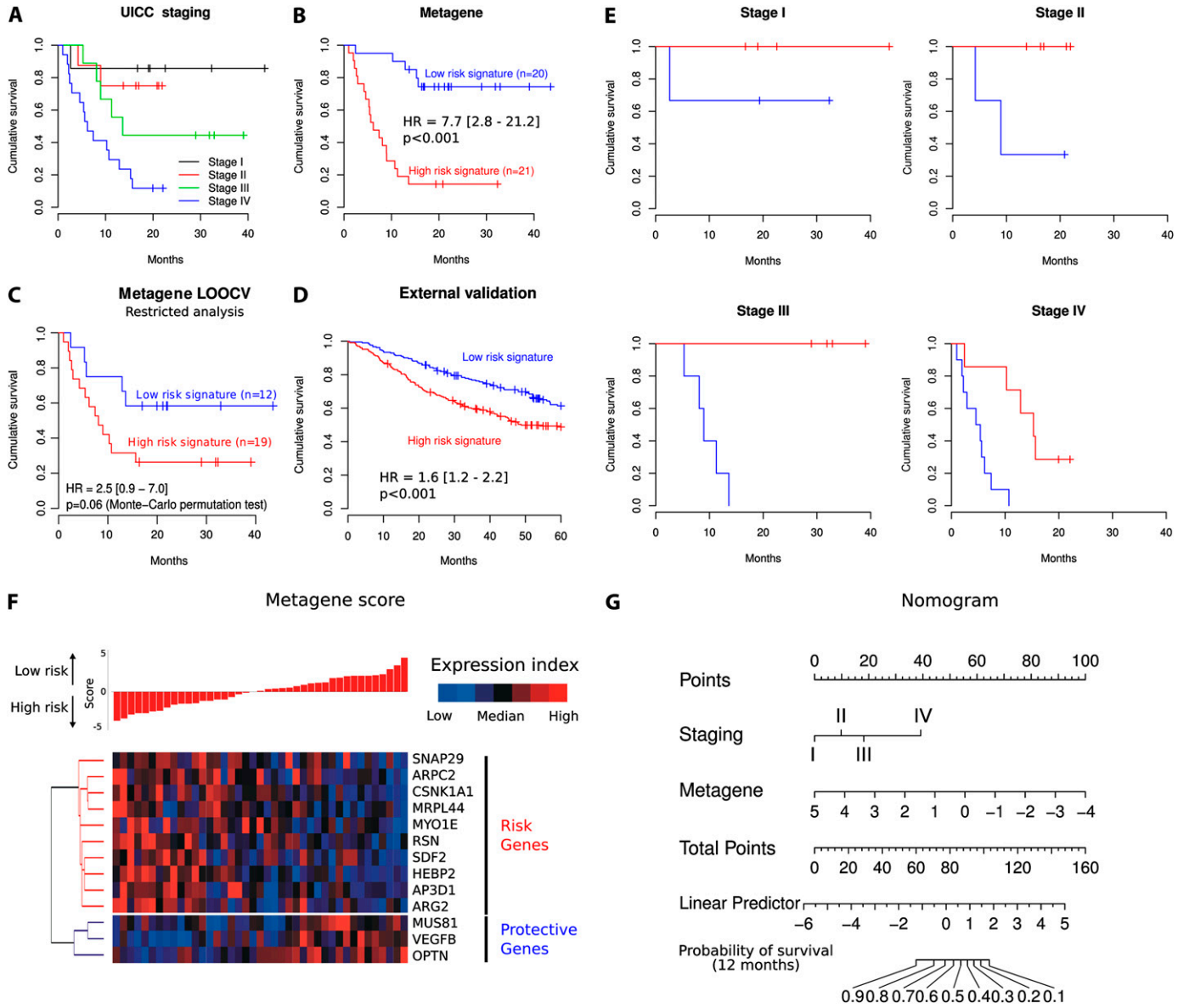

Figure 2. Survival analysis. (*A* and *B*) Kaplan-Meier estimates of survival according to the four International Union Against Cancer (UICC) stages and the risk estimated by the 13-gene metagene, respectively. (*C*) How the 13-gene metagene applies to our data once the analysis has been restricted to the bronchoscopic biopsies only. (*D*) Prognostic performance of the 13-gene metagene in an independent data set (11). (*E*) Kaplan-Meier estimates of survival according to the metagene scores in the different UICC stages. (*F*) Gene expression profiles of individual patients with non–small cell lung cancer (each column represents a patient) for the 13 prognostic genes included in the metagene. The magnitude of the metagene score is in relation to survival: the lower the score, the higher the risk of early death. (*G*) Nomogram summarizes the contribution of both staging and metagene on survival and can be used to predict survival manually. The upper scale (Points) is used to get the individual contribution of the two factors. The sum of both contributions must be reported in the lower scale (Total Points), which in turn is used to predict the probability of survival at 12 months. For example, an individual with Stage III and a metagene of 1 will accumulate approximately 60 points, which corresponds to an expected 12-month survival probability of 90%, whereas an individual with the same staging but a metagene of −2 will accumulate approximately 100 points and have a 12-month survival probability of 10%. HR = hazard ratio [95% confidence interval].

12-month survival. When combining both the UICC stage and the metagene, a significant gain of fit was obtained ($P < 0.001$; Table 3). Figure 2F displays the metagene score and the expression intensity of the 13 genes for each patient. The three protective genes included the mus81 endonuclease homolog (yeast) (*MUS81*), optineurin (*OPTN*), and *VEGFB* (*see* the Figures section of the online supplement).

A restricted analysis was done using only the specimen obtained from bronchoscopic biopsy. In this subanalysis, a 63-gene metagene gave the most accurate prediction of survival. When compared with the original 44-gene metagene, there was an overlap of 38 genes and each of the 13 genes from the original refined metagene was present in the new metagene. A leave-one-out cross-validation was performed to assess the prediction performance of a survival classifier based on the bronchoscopic biopsies alone. Each sample was sequentially removed from the initial data set and a metagene was built on each resampled data set, which was in turn used to predict the leftover sample. For each resampled data set, the gene selection procedure was done based on univariate Cox pro-

TABLE 3. COMPARISON OF THE PROGNOSTIC EFFICIENCY OF THE METAGENE COMPARED TO THE INTERNATIONAL UNION AGAINST CANCER STAGE

Model	Parameter	Hazard Ratio [95% CI]	*P* Value	Likelihood Ratio Test
1	UICC stages	2.4 [1.4–4.0]	7.9×10^{-4}	9.2×10^{-5}
2	Metagene scores	2.1 [1.6–2.9]	1.3×10^{-6}	6.7×10^{-8}
3	UICC stages +	2.9 [1.6–5.3]	4.5×10^{-4}	9.2×10^{-11}
	Metagene scores	2.2 [1.6–2.9]	8.2×10^{-7}	

Definition of abbreviations: CI = confidence interval; UICC = International Union against Cancer.

Three Cox proportional-hazards regression models were fitted, including UICC stage, metagene score, and both stage and metagene score. The hazard ratios, 95% CI, *P* values, and likelihood ratio tests obtained from these model fits are provided. There is a significant improvement of survival prediction when using the metagene information in combination with the UICC stage ($P < 0.001$).

portional hazards regression and refined by backward variable selection (28). A metagene was built by linear combination of the selected genes (first principal component of a supervised principal components analysis). The metagene was subsequently used for prediction of survival (low/high-risk, depending on the median of the metagene scores). The result of these predictions is depicted in Figure 2C. A distinct separation between the high-risk and the low-risk population is shown, although the statistical significance was only borderline ($P = 0.06$, Monte Carlo permutation test).

Nine duplicated samples were available for the prediction of survival. Over the 13 genes of the refined metagene, the intrain-dividual variability was found to be comparatively low (median CV = 2.3%) and a good prognostic repeatability was obtained (Figure E5B).

Validation of the Survival Signature

The 13-gene metagene was tested on the data set of director's challenge test consortium for the molecular classification of lung adenocarcinoma (11). When applied to this independent data set, the metagene (first principal component of a supervised principal components analysis) was a significant predictor of survival (hazard ratio = 1.6, 95% CI [1.2–2.2], $P < 0.001$) (Figure 2D). The prediction of survival by the metagene was independent of the staging because it remained significant after adjusting for the UICC stage.

The protective gene *VEGFB* was validated on a protein level by tissue microarray analysis (Figure 3). The Kaplan-Meier analysis showed a significant relationship between the staining intensity of VEGFB and survival ($P < 0.001$). This association was independent of the staging. A high *VEGFB* protein level was associated with longer survival independent of the UICC stage.

Effect of the Biopsies' Tumor Proportion on the Prediction Accuracy

The impact of tumor cell content in the biopsies was assessed in terms of diagnostic and prognostic accuracy (Figure 4). The prediction accuracy was dependent on the presence and proportion of tumor cells present in the biopsies (Kruskal-Wallis test: $P < 0.001$). The median diagnostic accuracy was 39% when no tumor cells were found in the biopsies, whereas it was 87% in case of at least 1% visible tumor cells. On the other hand, the prognostic accuracy of the metagene—as measured by the absolute value of the individual residual error—did not significantly differ with varying degree of tumor cell content (Kruskal-Wallis test: $P = 0.95$). The patients with no tumor cell content (n = 8) were all sampled by bronchoscopic biopsies and had the following staging distribution: two stage II, one stage III, and five stage IV. Similar findings were obtained from the initial 44-gene metagene (Kruskal-Wallis test: $P = 0.55$).

DISCUSSION

In this study, gene expression microarray technology was applied to biopsies obtained during the routine clinical work-up of patients with suspicion of early-stage or late-stage lung cancer. It allowed for molecular classification and prediction of survival complementary to histopathologic examination and UICC stage information. All previous gene expression microarray studies investigating NSCLC patient outcome used tumor biopsies from surgical resections or CT-guided biopsies, which mostly limited their applicability to resectable cases (i.e., early stages). Using a high-sensitivity gene expression microarray technology we could also include minute bronchoscopic biopsies for hybridization. As a result we were able to include a high proportion of stage III (22%) and stage IV patients (41%) in our study. Thus, in contrast to former studies the proposed approach is also suitable for patients with late-stage NSCLC who constitute the largest group in clinical practice.

Sixty-seven percent of genes that discriminated between SCC, AC, and control patients have already been described in the literature as being associated with lung cancer. This clearly documents the reliability of gene signatures from mixed cell biopsies.

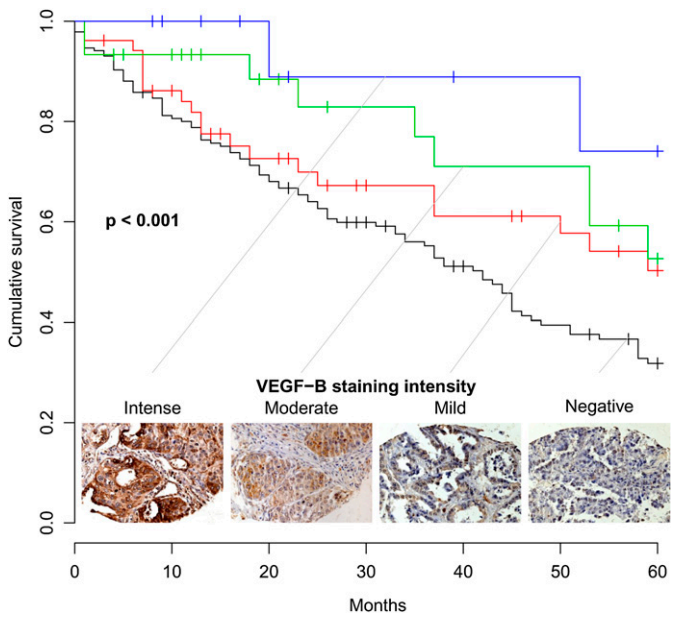

Figure 3. Validation of the prognostic value of vascular endothelial growth factor–β (VEGFB) staining intensity on tissue microarrays (TMA). Kaplan-Meier survival curves based on TMA-assessed protein expression levels are given. The staining categories were subdivided into intense (n = 13), moderate (n = 30), mild (n = 52), and negative (n = 188).

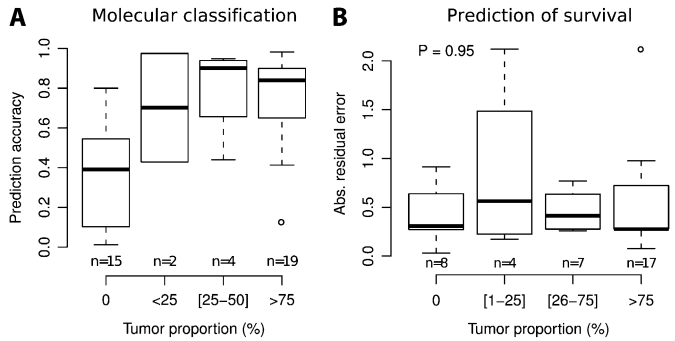

Figure 4. Importance of the tumor cell proportion in the classification accuracy and prediction of survival. (*A*) For a reliable classification accuracy, at least 1% and ideally more than 25% of visible tumor cells should be present in the biopsy. (*B*) Conversely, the prediction accuracy of survival was independent of the tumor cell content.

With 80% sensitivity and 89% specificity, the molecular classification was good. SCC typically presented high levels of genes involved in keratinization (29), overexpression of the aldo-keto reductase *AKR1B10* (30), and an up-regulation of several members of the calcium binding protein family (e.g., *S100A2*) (31). Moreover, the small proline-rich protein SPRR1B was identified as a marker for SCC (32), whereas in pulmonary AC a low level of SPRR family proteins was hypothesized to be associated with increased invasion into adjacent tissues (33). Interestingly, the S100 calcium-binding proteins and the SPRR proteins are located in the same chromosomal region (1q21) known to be highly rearranged in NSCLC (34). On the other hand, some surfactant proteins as well as NAPSA are used as diagnostic markers of AC of the lung. SFTPA and SFTPB are known to be tissue-specific markers expressed in normal and lung AC but not in SCC (35, 36), and NAPSA was shown to have an expression associated with the level of differentiation of lung AC (35).

With the aid of a metagene including 44 genes it was possible to accurately predict survival of our patients with NSCLC. Of these 44 genes, 13 genes were also associated with survival in at least one of four independent NSCLC data sets. With these 13 genes a refined metagene was constructed and successfully validated in another independent data set. The prognostic impact of *VEGF* family genes in NSCLC is well documented and usually associated with a bad prognosis (37, 38). In contrast, we found a favorable association between VEGFB and survival. This was confirmed at the protein level using tissue microarray on a cohort of 508 patients with NSCLC. Olofsson and colleagues showed that VEGFB can form heterodimers with VEGFA (39), a protein known as a key angiogenic player in cancer, which essentially acts via the receptor VEGFR-2 (40). They proposed that VEGFB/VEGFA heterodimerization could decrease VEGFA activity via VEGFR-2. Thus, VEGFB could play the role of a "natural" antagonist of VEGF signaling inhibiting tumor-promoting angiogenesis. As already described by others (41), we found that *ARG2* expression is associated with a bad prognosis. The arginases (ARG1 and ARG2) are required for cell cycle progression and they have the potential to sustain a high proliferation rate of tumor cells (42).

By cross-validation, we could assess the predictive value of signatures obtained from samples restricted to bronchoscopic biopsies only. Predictions from these signatures classify patients into high- or low-risk populations with a borderline significance.

The limitations of the current study include the limited number of patients. The lack of enrichment of tumor cells (e.g., by means of laser capture microdissection) is likely to result in smaller tumor-specific gene signatures. On the other hand, tumor signals can also spread away from the tumor into adjacent lung tissues and be picked up, a phenomenon described as "field cancerization." We could show that the biopsies should contain at least 1% (ideally $\geqslant$ 25%) of tumor cells to guarantee a reliable molecular classification. The prediction of survival, on the other hand, did not seem to be dependent on the proportion of tumor cells present in the biopsies. It appears that the host and/or tissue surrounding the tumor would carry sufficient and significant prognostic gene expression signals. These results support the hypothesis that the interplay between tumor and host and not the tumor alone is responsible for the outcome. Another limitation regards the smoking status and cumulative tobacco exposure, which was not assessed in this study and which might constitute a potential confounding role in the analysis.

In conclusion, gene expression signatures from biopsies taken during the initial work-up for NSCLC, including a significant proportion of minute bronchoscopic biopsies, can be used to establish a molecular diagnosis and to predict survival. We were able to identify a 13-gene metagene, which was successfully validated on an independent data set. This metagene was at least equivalent, and complementary, to the UICC stages for the prediction of survival. It proved to be particularly efficient for the identification of patients with a survival of less than 1 year independently of the UICC tumor stage information. Overall, the proposed strategy can be integrated into the daily clinical routine and is potentially useful to personalize treatment modalities in the future.

Conflict of Interest Statement: F.B. does not have a financial relationship with a commercial entity that has an interest in the subject of this manuscript. M.F. does not have a financial relationship with a commercial entity that has an interest in the subject of this manuscript. S.K. is a full-time employee of Novartis Pharma AG ($100,001 or more). M.S. is an employee of Novartis Pharma AG. M.P. served on an advisory board for Lily, Novartis, Roche, Sanofi, and AstraZeneca (up to $1,000). L.B. serves on the advisory board for Eli Lilly, AstraZeneca ($1,001–$5,000), received lecture fees from Abbott Molecular, Inc. ($5,001–$10,000), and received a sponsored grant from a noncommercial entity, Krebsliga beider Basel ($50,001–$100,000). S.S. does not have a financial relationship with a commercial entity that has an interest in the subject of this manuscript. E.M. does not have a financial relationship with a commercial entity that has an interest in the subject of this manuscript. W.B. does not have a financial relationship with a commercial entity that has an interest in the subject of this manuscript. M.B. does not have a financial relationship with a commercial entity that has an interest in the subject of this manuscript. J.K. is a full-time employee of Sanofi-Aventis and Novartis. M.T. does not have a financial relationship with a commercial entity that has an interest in the subject of this manuscript. M.H.B. has served on an advisory board for GlaxoSmithKline, Pfizer, Novartis, and Boehringer ($1,001–$5,000), lecture fees paid by GSK (up to $1,000), industry-sponsored grants from AZ ($10,001–$50,000), and Actelion and GSK ($5,001–$10,000).

Acknowledgment: The authors thank Isabelle Charmont, Christian Fuhrimann, and Edgar Baer, who performed RNA isolation and NovaChip hybridization; Daniel Wahl, who prepared the data submission to Gene Expression Omnibus; and Martin Früh for his constructive input reviewing the manuscript.

References

1. Scagliotti GV, Parikh P, von Pawel J, Biesma B, Vansteenkiste J, Manegold C, Serwatowski P, Gatzemeier U, Digumarti R, Zukin M, *et al.* Phase III study comparing cisplatin plus gemcitabine with cisplatin plus pemetrexed in chemotherapy-naive patients with advanced-stage non-small-cell lung cancer. *J Clin Oncol* 2008;26:3543–3551.
2. Shigematsu H, Lin L, Takahashi T, Nomura M, Suzuki M, Wistuba II, Fong KM, Lee H, Toyooka S, Shimizu N, *et al.* Clinical and biological features associated with epidermal growth factor receptor gene mutations in lung cancers. *J Natl Cancer Inst* 2005;97:339–346.
3. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, *et al.* Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci USA* 2001;98:13790–13795.
4. Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, van de Rijn M, Rosen GD, Perou CM, Whyte

RI, *et al.* Diversity of gene expression in adenocarcinoma of the lung. *Proc Natl Acad Sci USA* 2001;98:13784–13789.
5. Beer DG, Kardia SLR, Huang C, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, *et al.* Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med* 2002;8:816–824.
6. Tsao M, Sakurada A, Cutz J, Zhu C, Kamel-Reid S, Squire J, Lorimer I, Zhang T, Liu N, Daneshmand M, *et al.* Erlotinib in lung cancer—molecular and clinical predictors of outcome. *N Engl J Med* 2005;353: 133–144.
7. Raponi M, Zhang Y, Yu J, Chen G, Lee G, Taylor JMG, Macdonald J, Thomas D, Moskaluk C, Wang Y, *et al.* Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. *Cancer Res* 2006;66:7466–7472.
8. Tomida S, Koshikawa K, Yatabe Y, Harano T, Ogura N, Mitsudomi T, Some M, Yanagisawa K, Takahashi T, Osada H, *et al.* Gene expression-based, individualized outcome prediction for surgically treated lung cancer patients. *Oncogene* 2004;23:5360–5370.
9. Lu Y, Lemon W, Liu P, Yi Y, Morrison C, Yang P, Sun Z, Szoke J, Gerald WL, Watson M, *et al.* A gene expression signature predicts survival of patients with stage I non-small cell lung cancer. *PLoS Med* 2006;3:e467.
10. Chen H, Yu S, Chen C, Chang G, Chen C, Yuan A, Cheng C, Wang C, Terng H, Kao S, *et al.* A five-gene signature and clinical outcome in non-small-cell lung cancer. *N Engl J Med* 2007;356:11–20.
11. Shedden K, Taylor JMG, Enkemann SA, Tsao M, Yeatman TJ, Gerald WL, Eschrich S, Jurisica I, Giordano TJ, Misek DE, *et al.* Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat Med* 2008;14:822–827.
12. Guo NL, Wan Y, Tosun K, Lin H, Msiska Z, Flynn DC, Remick SC, Vallyathan V, Dowlati A, Shi X, *et al.* Confirmation of gene expression-based prediction of survival in non-small cell lung cancer. *Clin Cancer Res* 2008;14:8213–8220.
13. Borczuk AC, Shah L, Pearson GDN, Walter KL, Wang L, Austin JHM, Friedman RA, Powell CA. Molecular signatures in biopsy specimens of lung cancer. *Am J Respir Crit Care Med* 2004;170:167–174.
14. Virtanen C, Ishikawa Y, Honjoh D, Kimura M, Shimane M, Miyoshi T, Nomura H, Jones MH. Integrated classification of lung tumors and cell lines by expression profiling. *Proc Natl Acad Sci USA* 2002;99: 12357–12362.
15. Imperatori A, Harrison RN, Leitch DN, Rovera F, Lepore G, Dionigi G, Sutton P, Dominioni L. Lung cancer in Teesside (UK) and Varese (Italy): a comparison of management and survival. *Thorax* 2006;61:232–239.
16. Spira A, Beane JE, Shah V, Steiling K, Liu G, Schembri F, Gilman S, Dumas Y, Calner P, Sebastiani P, *et al.* Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nat Med* 2007;13:361–366.
17. British Thoracic Society Bronchoscopy Guidelines Committee. British Thoracic Society guidelines on diagnostic flexible bronchoscopy. *Thorax* 2001;56 suppl 1:i1–21.
18. Ettinger DS, Akerley W, Bepler G, Chang A, Cheney RT, Chirieac LR, D'Amico TA, Demmy TL, Feigenberg SJ, Figlin RA, *et al.* Non-small cell lung cancer. *J Natl Compr Canc Netw* 2008;6:228–269.
19. Marrer E, Baty F, Kehren J, Chibout S, Brutsche M. Past, present and future of gene expression-tailored therapy for lung cancer. *Per Med* 2006;3:165–175.
20. Baty F, Buess M, Kaiser S, Facompré M, Bubendorf L, Schumacher M, Budach W, Kehren J, Brutsche M. Prediction of major NSCLC tumor classes and clinical outcome by functional genomics of small bronchoscopic tumor biopsies: B7–03. *J Thorac Oncol* 2007;2:S355.
21. Neuschäfer D, Budach W, Wanke C, Chibout SD. Evanescent resonator chips: a universal platform with superior sensitivity for fluorescence-based microarrays. *Biosens Bioelectron* 2003;18:489–497.
22. Culhane AC, Perrière G, Considine EC, Cotter TG, Higgins DG. Between-group analysis of microarray data. *Bioinformatics* 2002;18: 1600–1608.
23. Baty F, Facompré M, Wiegand J, Schwager J, Brutsche MH. Analysis with respect to instrumental variables for the exploration of microarray data structures. *BMC Bioinformatics* 2006;7:422.
24. Trevino V, Falciani F. GALGO: an R package for multivariate variable selection using genetic algorithms. *Bioinformatics* 2006;22:1154–1156.
25. Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol* 2004;2:E108.
26. Bild AH, Potti A, Nevins JR. Linking oncogenic pathways with therapeutic opportunities. *Nat Rev Cancer* 2006;6:735–741.
27. Kononen J, Bubendorf L, Kallioniemi A, Bärlund M, Schraml P, Leighton S, Torhorst J, Mihatsch MJ, Sauter G, Kallioniemi OP. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med* 1998;4:844–847.
28. Baty F, Bihl MP, Perrière G, Culhane AC, Brutsche MH. Optimized between-group classification: a new jackknife-based gene selection procedure for genome-wide expression data. *BMC Bioinformatics* 2005;6:239.
29. Proksch E, Fölster-Holst R, Jensen J. Skin barrier function, epidermal proliferation and differentiation in eczema. *J Dermatol Sci* 2006;43: 159–169.
30. Fukumoto S, Yamauchi N, Moriguchi H, Hippo Y, Watanabe A, Shibahara J, Taniguchi H, Ishikawa S, Ito H, Yamamoto S, *et al.* Overexpression of the aldo-keto reductase family protein AKR1B10 is highly correlated with smokers' non-small cell lung carcinomas. *Clin Cancer Res* 2005;11:1776–1785.
31. Zech VFE, Dlaska M, Tzankov A, Hilbe W. Prognostic and diagnostic relevance of hnRNP A2/B1, hnRNP B1 and S100 A2 in non-small cell lung cancer. *Cancer Detect Prev* 2006;30:395–402.
32. Hu R, Wu R, Deng J, Lau D. A small proline-rich protein, spr1: specific marker for squamous lung carcinoma. *Lung Cancer* 1998;20:25–30.
33. Nacht M, Dracheva T, Gao Y, Fujii T, Chen Y, Player A, Akmaev V, Cook B, Dufault M, Zhang M, *et al.* Molecular characteristics of non-small cell lung cancer. *Proc Natl Acad Sci USA* 2001;98:15203–15208.
34. Whang-Peng J, Knutsen T, Gazdar A, Steinberg SM, Oie H, Linnoila I, Mulshine J, Nau M, Minna JD. Nonrandom structural and numerical chromosome changes in non-small-cell lung cancer. *Genes Chromosomes Cancer* 1991;3:168–188.
35. Khoor A, Whitsett JA, Stahlman MT, Halter SA. Expression of surfactant protein B precursor and surfactant protein B mRNA in adenocarcinoma of the lung. *Mod Pathol* 1997;10:62–67.
36. Zamecnik J, Kodet R. Value of thyroid transcription factor-1 and surfactant apoprotein A in the differential diagnosis of pulmonary carcinomas: a study of 109 cases. *Virchows Arch* 2002;440:353–361.
37. Bremnes RM, Camps C, Sirera R. Angiogenesis in non-small cell lung cancer: the prognostic impact of neoangiogenesis and the cytokines VEGF and bFGF in tumours and blood. *Lung Cancer* 2006;51:143–158.
38. Sandler A, Gray R, Perry MC, Brahmer J, Schiller JH, Dowlati A, Lilenbaum R, Johnson DH. Paclitaxel-carboplatin alone or with bevacizumab for non-small-cell lung cancer. *N Engl J Med* 2006; 355:2542–2550.
39. Olofsson B, Korpelainen E, Pepper MS, Mandriota SJ, Aase K, Kumar V, Gunji Y, Jeltsch MM, Shibuya M, Alitalo K, *et al.* Vascular endothelial growth factor B (VEGF-B) binds to VEGF receptor-1 and regulates plasminogen activator activity in endothelial cells. *Proc Natl Acad Sci USA* 1998;95:11709–11714.
40. Takahashi H, Shibuya M. The vascular endothelial growth factor (VEGF)/VEGF receptor system and its role under physiological and pathological conditions. *Clin Sci* 2005;109:227–241.
41. Süer Gökmen S, Yörük Y, Cakir E, Yorulmaz F, Gülen S. Arginase and ornithine, as markers in human non-small cell lung carcinoma. *Cancer Biochem Biophys* 1999;17:125–131.
42. Rodriguez PC, Quiceno DG, Zabaleta J, Ortiz B, Zea AH, Piazuelo MB, Delgado A, Correa P, Brayer J, Sotomayor EM, *et al.* Arginase I production in the tumor microenvironment by mature myeloid cells inhibits T-cell receptor expression and antigen-specific T-cell responses. *Cancer Res* 2004;64:5839–5849.