The lung-specific proteome defined by integration of transcriptomics and antibody-based profiling

Cecilia Lindskog,* Linn Fagerberg,† Björn Hallström,† Karolina Edlund,‡ Birte Hellwig,§ Jörg Rahnenführer,‡ Caroline Kampf,* Mathias Uhlén,† Fredrik Pontén,*,[1,2] and Patrick Micke*,[1]

*Rudbeck Laboratory, Science for Life Laboratory, Uppsala University, Uppsala, Sweden; †Science for Life Laboratory, Kungliga Tekniska Högskolan (KTH) Royal Institute of Technology, Stockholm, Sweden; ‡Leibniz Research Centre for Working Environment and Human Factors (IfADo) and §Department of Statistics, Dortmund Technical University, Dortmund, Germany

ABSTRACT **The combined action of multiple cell types is essential for the physiological function of the lung, and increased awareness of the molecular constituents characterizing each cell type is likely to advance the understanding of lung biology and disease. In the current study, we used genome-wide RNA sequencing of normal lung parenchyma and 26 additional tissue types, combined with antibody-based protein profiling, to localize the expression to specific cell types. Altogether, 221 genes were found to be elevated in the lung compared with their expression in other analyzed tissues. Among the gene products were several well-known markers, but also several proteins previously not described in the context of the lung. To link the lung-specific molecular repertoire to human disease, survival associations of pneumocyte-specific genes were assessed by using transcriptomics data from 7 non-small-cell lung cancer (NSCLC) cohorts. Transcript levels of 10 genes (*SFTPB*, *SFTPC*, *SFTPD*, *SLC34A2*, *LAMP3*, *CACNA2D2*, *AGER*, *EMP2*, *NKX2-1*, and *NAPSA*) were significantly associated with survival in the adenocarcinoma subgroup, thus qualifying as promising biomarker candidates. In summary, based on an integrated omics approach, we identified genes with elevated expression in lung and localized corresponding protein expression to different cell types. As biomarker candidates, these proteins may represent intriguing starting points for further exploration in health and disease.—Lindskog, C., Fagerberg, L., Hallström, B., Edlund, K., Hellwig, B., Rahnenführer, J., Kampf, C., Uhlén, M., Pontén, F., Micke, P. The lung-specific proteome defined by integration of transcriptomics and antibody-based profiling. *FASEB J.* 28, 5184–5196 (2014). www.fasebj.org**

Key Words: patients · prognosis · expression analysis · in situ detection

THE DELICATE STRUCTURE of lung tissue is shaped by several different cell types that together are indispensable for the physiological functions of the lung. Detailed knowledge of the molecular repertoire of each constituent is essential in understanding the cellular interactions under normal and pathological conditions. In pulmonary disease, gene expression profiling has been used widely to unravel the biology of immune, infectious, and malignant conditions (1). Thus far, primarily microarray-based techniques have been applied—for instance, to subclassify lung cancer entities, predict therapy response, and identify underlying mechanisms of tumor development (2–4).

In contrast to array-based methods, next-generation transcriptome sequencing (RNA-Seq) provides a numerical description of absolute transcript levels, down to an average of a few molecules per cell (5). Although this technique provides outstanding analytical power to determine mRNA expression levels, the biological functions of a cell are principally carried out by the translated protein repertoire, and gene expression profiling can thus offer only indications of cellular mechanisms. Moreover, as human tissue is notoriously heterogeneous, comprising different cell types in variable proportions, detected transcript levels represent average values of the analyzed cell mixture and thus only provide hints with regard to cell type specificity. By using antibody-based profiling, genes detected by transcriptomics studies can be further explored on the protein level at single-cell resolution *in situ* to place findings in the proper context of tissue microenvironment and subcellular localization.

Based on a previous transcriptomics analysis, including 27 different human organ sites (6), the purpose of

Abbreviations: FPKM, fragments per kilobase of exon model per million mapped reads; GO, gene ontology; NSCLC, non-small-cell lung cancer; TMA, tissue microarray

[1] These authors contributed equally to this work.

[2] Correspondence: Science for Life Laboratory, Department of Immunology, Genetics, and Pathology, Uppsala University, Dag Hammarskjölds väg 20, 751 85 Uppsala, Sweden. E-mail: fredrik.ponten@igp.uu.se

doi: 10.1096/fj.14-254862

This article includes supplemental data. Please visit *http://www.fasebj.org* to obtain this information.

0892-6638/14/0028-5184 © FASEB

the current study was to determine the lung-specific transcriptome in comparison to 26 other tissue types based on RNA-Seq. The analysis was combined with antibody-based protein profiling using antibody resources generated within the Human Protein Atlas program (http://www.proteinatlas.org). The identification of lung-specific expression may serve as a starting point for further disease-specific research, exemplified in this study by the identification of pneumocyte-specific gene expression with a significant association to overall survival in non-small-cell lung cancer (NSCLC).

MATERIALS AND METHODS

Sample characteristics

The use of human tissue samples was approved by the Uppsala Ethics Review Board (2011/473). Fresh-frozen tissues from 27 different, normal body sites were included, as described previously (6). The samples were embedded in optimal cutting temperature medium and stored at −80°C. Frozen 4 μm sections from all tissues were stained with hematoxylin and eosin and examined by a pathologist (F.P.), to ensure proper morphology. The surgical material used for transcript profiling of human lung included histologically normal tissue from 1 lung emphysema (individual 1), 1 lung carcinoid (individual 2), and 3 samples of squamous cell carcinoma (individuals 3–5). The corresponding histology of the 5 cases is shown in Supplemental Fig. S1. The average proportion of different cell types in the samples was estimated to be 36% pneumocytes, 28% endothelial cells, 11% macrophages, 6% bronchial epithelium, and 19% other cell types (inflammatory cells, smooth muscle cells, and fibroblasts). Four of the 5 samples included a significant number of bronchial epithelial cells, with the exception of sample 3.

Transcript profiling (RNA-Seq)

Three sections (10 μm) were cut from each frozen tissue block and homogenized before extraction of total RNA with an RNeasy Mini Kit (Qiagen, Hilden, Germany), according to the manufacturer's instructions. Extracted RNA samples were analyzed with either an Experion automated electrophoresis system (Bio-Rad Laboratories, Hercules, CA, USA) with the standard-sensitivity RNA chip, or the Agilent 2100 Bioanalyzer system with the RNA 6000 Nano LabChip Kit (Agilent Biotechnologies, Palo Alto, CA, USA). Only samples of high-quality RNA (RNA integrity number $\geq$7.5) were used for mRNA sequencing, performed on Illumina HiSeq2000/2500 machines (Illumina, San Diego, CA, USA), using the standard RNA-Seq protocol with a read length of 2 × 100 bases.

Data analysis

Raw reads from 20,050 transcripts were trimmed for low-quality ends with the Sickle software (7), using a phred quality threshold of 20. Processed reads were mapped to the GRCh37 version of the human genome with Tophat v2.0.3 (8). Potential PCR duplicates were eliminated applying the MarkDuplicates module of Picard 1.77 (9). To obtain quantification scores for all human genes, fragments per kilobase of exon model per million mapped reads (FPKM) values were calculated by using gene models from Ensembl build 69 with Cufflinks 2.0.2 (8), which corrects for transcript length and the total number of mapped reads from the library, compensating for different read depths for different samples. All data were analyzed with R Statistical Environment, and a network analysis was performed with Cytoscape 3.0 (10). For the analyses performed in this study where a $\log_2$ scale of the data was used, pseudocounts of +1 were added to the data set.

Specificity classification

The average FPKM value of all individual samples for each tissue was used to estimate the gene expression level. A cutoff of 1 FPKM was used as the detection limit, roughly corresponding to 1 mRNA per average cell in the sample (11). Each of the 20,050 gene transcripts was classified into 1 of 7 categories based on the FPKM levels: not detected (<1 FPKM in lung tissue); highly lung enriched (50-fold higher FPKM level in lung compared with the other 26 tissues); moderately lung enriched (5-fold higher FPKM level in lung compared with the other tissues); group enriched (5-fold higher average FPKM level in a group of 2–7 tissues including lung, compared with the other tissues); expressed in all (detected in all 27 tissues); lung enhanced (5-fold higher FPKM level in lung compared with the average FPKM value of all 27 tissues); and mixed (genes expressed in 1–26 tissues and in none of the other categories). The lung-specific score was defined as the lung FPKM divided by the maximum FPKM in any of the other 26 tissues.

Gene ontology (GO) analysis

A GO-based analysis (12) was performed with the GOrilla tool (13), to determine overrepresented GO categories in the gene set of elevated lung genes. For the cellular component analysis, the GOSlim GO analysis associations were used to determine whether genes encoded extracellular, intracellular, or membrane-bound proteins. The number of genes for each term was counted, allowing a gene to be associated with >1 term. A list of all the genes analyzed in this study was used as the background list in GOrilla.

Antibody-based profiling

Tissue microarrays (TMAs) were generated as described previously (14), containing 1 mm cores of 44 different normal tissues in triplicate. Samples were received from the Department of Pathology, Uppsala University Hospital, and approved by the local Research Ethics Committee (Uppsala, Sweden; Ups 02–577). Automated immunohistochemistry was performed (14). Tissue incubated with PBS instead of primary antibody served as the negative controls. Immunohistochemically stained and mounted slides were scanned with a ScanScope XT Slide Scanner (Aperio Technologies, Vista, CA, USA), to generate high-resolution digital images, followed by annotation of intensity and fraction of positive cells defined in the different tissues.

Survival analysis

Seven publically available NSCLC data sets, altogether comprising 1252 patients analyzed on Affymetrix GeneChip HG U133A or U133 Plus 2.0 arrays (Affymetrix, Santa Clara, CA, USA), with accompanying information on overall survival, were downloaded: GSE14814 (15), GSE19188 (16), GSE31210 (17), GSE3141 (18), GSE4573 (19), GSE37745 (3), and Shedden *et al.* (2) Two datasets (GSE31210 and Shedden *et al.*, ref. 2) included only samples from patients with adenocarcinoma, whereas 1 other dataset (GSE4573) included only squamous cell carcinomas, leading to 6 and 5 datasets being included in

the separate analysis of the adenocarcinoma and squamous cell carcinoma subtypes, respectively. A robust multiarray average (RMA) was used for normalization (20). Univariate Cox models were used to analyze the association of the expression of a specific gene with overall survival in each study separately. To combine single estimates of different studies into 1 pooled overall estimate, we applied random-effects meta-analysis models based on parameter estimates of log hazard ratios and their corresponding standard errors, using the R package meta (http://cran.r-project.org/web/packages/meta/) with default settings. *P* values of the meta-analyses were adjusted for multiple testing using the method of Benjamini and Hochberg (21), which controls the false-discovery rate (FDR).

Data availability

All data (FPKM values for all samples) are available for download without any restrictions (http://www.proteinatlas.org/about/download/). The primary data (reads) are available through the Array Express Archive (http://www.ebi.ac.uk/arrayexpress/) under the accession number E-MTAB-1733. Transcript profiling data (FPKM values) for each gene in each cell and tissue type are also available at the Human Protein Atlas website (http://www.proteinatlas.org), together with all annotation data for the different antibodies.

RESULTS

Transcriptomic analysis of lung tissue

A transcriptomic analysis was performed with RNA-Seq data based on fresh-frozen tissue from 27 different tissue types, including lung samples from 5 individuals, as described earlier (6). Altogether, 95 samples were analyzed, and the present investigation focused on genes expressed in the lung. The transcriptome of each sample was quantified to determine normalized mRNA levels, calculated as FPKM (8). We used the cutoff of 1 FPKM, roughly corresponding to 1 mRNA/average cell in the sample (11). The mRNA expression in lung ranged from 8409 down to the statistical cutoff of 0.1, yielding a dynamic range of $\sim 10^5$ between the highest and lowest expressed genes within lung tissue. Among the 30 genes with highest expression levels in the lung are many genes with housekeeping functions, such as those encoding mitochondrial and ribosomal proteins. The number of genes with an expression value of >1 FPKM varied from 13,572 genes in individual 3 to 13,912 genes in individual 2. Thus, $\sim$70% of all putative protein-coding genes (n=20,050) were found to be expressed in human lung.

The variance between lung samples from different individuals was analyzed with pairwise Spearman correlations, analyzing the expression level of all protein coding genes. The Spearman correlations between 2 different lung samples ranged from 0.96 to 0.98, indicating low interindividual variations and low biological variance in the genome-wide expression pattern across lung samples.

Classification of the genes expressed in lung

Based on the transcriptomic data, the human protein coding genome was classified into the different categories illustrated in **Fig. 1*A***: genes not detected in the lung (30.5%); the largest group of expressed genes (46%), representing those expressed in all 27 tissue types; a mixed group of genes (22.4%) expressed in several, but not all, tissue types, with no enriched expression in the lung; and finally, genes with elevated expression in lung (1.1%). Of the genes with elevated expression in lung, only 6 were highly lung enriched, with a 50-fold higher FPKM level in lung compared with all other tissues (**Table 1**). In addition, 14 genes were moderately lung enriched with a 5-fold higher FPKM level in lung compared with levels in all other tissues (Table 1). Ninety-six genes were group enriched, with a 5-fold higher average FPKM level in a group of 2–7 tissues including lung, compared with all other tissues (Supplemental Table S1), and 105 genes were lung enhanced (*i.e.*, genes with an expression in the lung $\geq$5-fold higher than the average expression in all 27 tissues; Supplemental Table S2).

An in-depth analysis of the expression levels of each gene enabled calculation of the relative fraction of the mRNA pool for each of the categories, as displayed in Fig. 1*B*. The analysis showed that 82% of the mRNA molecules in the lung corresponded to the housekeeping genes expressed in all tissues, and only 8% of the mRNA pool corresponded to genes categorized as elevated in the lung. To illustrate the relation of lung tissue to other tissue types, a network plot was generated, displaying the number of commonly expressed genes between different tissues (Fig. 1*C*). The lung did not show a specific pattern of shared group-enriched genes with any of the other tissue types except for testis (33 genes). Interestingly, a GO-based analysis of these shared genes showed an overrepresentation of genes related to cilium function and movement.

An analysis of all the 221 genes elevated in lung was well in line with the function of the lung. Of the 6 highly enriched genes, 5 are surfactant proteins with the function to form an air–liquid interface in the alveoli. Also, among the moderately enriched genes, several surfactant genes were found. A GO-based analysis of the genes elevated in lung identified an overrepresentation of genes associated with endocytosis (12 genes), respiratory gaseous exchange (5 genes), cilium movement (4 genes), and surfactant homeostasis (3 genes). Approximately half (45%) of the gene products were located in the extracellular space, whereas 32% were found in the intracellular compartment, and 23% were part of the membrane region.

Antibody-based profiling of the lung-specific genes

The elevated genes identified by the transcriptomics analysis were further investigated by using antibody-based protein profiling as part of the Human Protein Atlas program, including immunohistochemistry-based

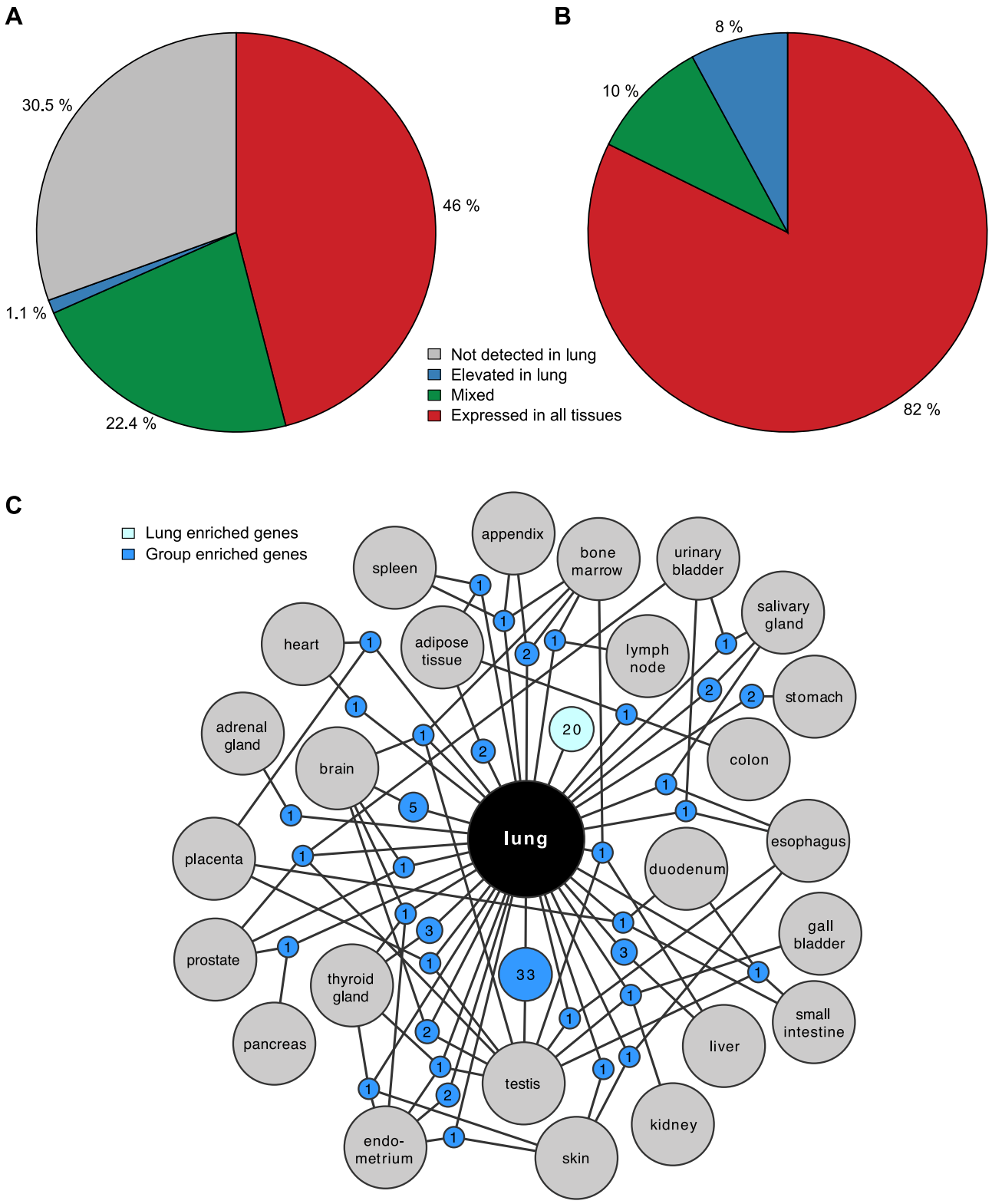

Figure 1. Classification of all human protein-coding genes. *A*) Distribution of all 20,050 genes into 4 different categories based on the transcript abundance, as well as the number of detected tissues. *B*) Distribution of the fraction of expressed mRNA molecules (*i.e.*, the sum of all FPKM values for each of the categories). *C*) Network plot of the lung- and group-enriched genes. Blue circles represent the number of enriched genes shared between a particular combination of tissues (gray circles).

TMA analysis of 3 cores of alveolar lung tissue and 3 cores of bronchial epithelium. For the 20 lung-enriched genes (Table 1), a high concordance between RNA-Seq data and immunohistochemistry data was observed for a majority (76%) of the 17 genes with available antibodies, essentially displaying high expression in lung tissue and negative or only low expression in other investigated tissue types.

TABLE 1. *Lung-enriched proteins*

Gene	Description	mRNA level (FPKM)	Lung-specific score	HPA (Abs)	Concordance mRNA/IHC
SFTPA1	Surfactant protein A1	4866	1206	4	Yes
SFTPB	Surfactant protein B	2744	726	2	Yes
SFTPC	Surfactant protein C	8409	312	2	Yes
SFTPA2	Surfactant protein A2	6139	275	4	Yes
SCGB3A2	Secretoglobin, family 3A, member 2	465	129	0	Yes
AGER	Advanced glycosylation receptor	1064	60	1	Yes
SFTPD	Surfactant protein D	461	44	2	Yes
ROS1	c-Ros oncogene 1	23	41	1	No
SCGB1A1	Secretoglobin (uteroglobin)	1803	39	2	Yes
MS4A15	Membrane-spanning, member 15	51	23	0	NA
RTKN2	Rhotekin 2	92	14	2	No
SLC34A2	Solute carrier family 34	619	14	1	Yes
NAPSA	Napsin A aspartic peptidase	730	12	4	Yes
SFTA2	Surfactant associated 2	242	7	0	NA
LRRN4	Leucine-rich repeat neuronal 4	9	7	1	No
LAMP3	Lysosomal-assoc. membrane protein 3	192	6	2	Yes
CACNA2D2	Calcium channel, α/δ subunit 2	48	6	1	Yes
FAM92B	Family with similarity 92, member B	5	6	1	Yes
CCDC17	Coiled-coil domain containing 17	10	6	1	Yes
LDLRAD1	Low-density lipoprotein receptor	9	5	1	No

List is sorted according to lung-specific score, which is the FPKM value in lung tissue divided by the maximum FPKM in all other tissues. HPA indicates the number of validated antibodies (Abs) publicly available on the Human Protein Atlas. Concordance mRNA/IHC shows the concordance between mRNA data and immunohistochemistry data (*i.e.*, the antibodies that showed elevated expression in lung and also at the protein level).

Antibody-based profiling allows further localization of protein expression to the various cell types present in lung tissue. Of the 221 genes defined as elevated in the lung, an in-depth immunohistochemistry analysis identified proteins with high immunoreactivity in the lung as compared to cell types present in the other 43 investigated tissues. The lung-specific expression was categorized into different cell types within the lung, such as pneumocytes and ciliated, mucus-secreting, endothelial, and immune cells (**Table 2**).

Analysis of pneumocyte-specific proteins

Alveolar pneumocytes signify lung tissue and are necessary for the maintenance of surfactant and for normal respiration. Examples of 13 proteins with specific expression in pneumocytes are displayed in **Fig. 2**. Five of them are surfactant proteins, expressed in type II pneumocytes (SFTPA1, SFTPA2, SFTPB, SFTPC, and SFTPD). Surfactants act by lowering the surface tension, crucial for gaseous exchange between air and blood. Moreover, surfactant proteins are involved in defense against microbial invasion. Another example is SLC34A2, also expressed in type II pneumocytes. This protein may be involved in synthesis of surfactant, and mutations in the corresponding gene can cause pulmonary alveolar microlithiasis (22). NAPSA, a secreted protein expressed in type II pneumocytes, is also associated with surfactant processing. Furthermore, because of its lung specificity, NAPSA is used as a marker in diagnostic pathology to determine tumor origin (23). Similarly, the transcription factor NKX2-1, better known as TTF1, is commonly applied as a diagnostic marker for the lung lineage. LAMP3 plays a role in dendritic cell function and in adaptive immunity and has been suggested to play a role in processing of the ABCA3 protein, a lipid transporter required for surfactant biogenesis (24). In lung tissue, LAMP3 expression is restricted to type II pneumocytes. CACNA2D2 expressed in type II pneumocytes is involved in calcium channel regulation and has been detected at high levels in lung by Northern blot analysis (25). Moreover, CACNA2D2 is part of a tumor-suppressor gene cluster and has been implicated in different studies on NSCLC (26). AGER is a cell surface receptor, and its ligand AGE is involved in several functions in blood vessels, including endothelial permeability, NO signaling, and endothelial progenitor cell function. Besides AGE, the receptor interacts with molecules implicated in homeostasis, development, and inflammation (27). In lung, a clear expression is evident in type I pneumocytes, accompanied with positivity in endothelial cells. CAV1 is a scaffolding protein and an essential part of caveolae plasma membranes. CAV1 regulates critical cell functions, including proliferation, apoptosis, cell differentiation, and transcytosis *via* diverse signaling pathways. Moreover, CAV1 has been suggested to be involved in lung development, lung cancer, pathogenesis of lung injury, and defense against infections (28). In lung tissue, CAV1 expression is mainly found in type I pneumocytes and endothelial cells. EMP2, with evidence of existence only at transcript level, seems to reside at the plasma membrane within lipid raft

TABLE 2. *Proteins selectively expressed in subsets of cells in lung tissue, with information regarding mRNA level, immunohistochemistry-based expression pattern, and antibody used*

Gene	Description	RNA category	mRNA level (FPKM)	Lung-specific score	Subcellular localization	HPA Ab example
Pneumocyte-specific proteins						
SFTPA1	Surfactant protein A1	High	4866	1206	Cyt, sec	HPA049368
SFTPA2	Surfactant protein A2	High	6139	275	Cyt, sec	CAB002439
SFTPB	Surfactant protein B	High	2744	726	Cyt	CAB002440
SFTPC	Surfactant protein C	High	8409	312	Cyt	HPA010928
SFTPD	Surfactant protein D	Mod	461	44	Cyt, sec	HPA044582
SLC34A2	Solute carrier family 34	Mod	619	14	Cyt, sec	HPA037989
NAPSA	Napsin A aspartic peptidase	Mod	730	12	Cyt	HPA045280
NKX2-1	NK2 homeobox 1	Group	55	1	Nucl	CAB000078
LAMP3	Lysosomal-associated membrane protein 3	Mod	192	6	Cyt	CAB025133
CACNA2D2	Calcium channel, voltage-dependent, α-2/δ subunit 2	Mod	48	6	Cyt	HPA034771
AGER	Advanced glycosylation end product-specific receptor	High	1064	60	Mem	CAB011682
CAV1	Caveolin 1, caveolae protein, 22kDa	Enh	610	1	Mem	CAB003791
EMP2	Epithelial membrane protein 2	Enh	243	2	Mem	HPA014711
Proteins specific for ciliated cells						
C1ORF87	Chromosome 1 open reading frame 87	Group	4	0.4	Mem	HPA031366
C9ORF17	Chromosome 9 open reading frame 171	Group	3	0.2	Mem	HPA021329
CCDC171	Coiled-coil domain containing 17	Mod	10	6	Mem	HPA028338
CCDC19	Coiled-coil domain containing 19	Group	6	0.3	Mem	HPA043618
CCDC37	Coiled-coil domain containing 37	Group	7	0.3	Mem	HPA046354
CCDC42B	Coiled-coil domain containing 42B	Enh	3	1	Mem	HPA048539
DYDC2	DPY30 domain containing 2	Group	6	1	Cyt, mem	HPA038006
FAM92B	Family with sequence similarity 92, member B	Mod	5	6	Mem	HPA041022
FAM183A	Family with sequence similarity 183, member A	Enh	17	2	Mem	HPA043382
DNAH5	Dynein, axonemal, heavy chain 5	Enh	2	2	Mem	HPA037470
DNAI1	Dynein, axonemal, intermediate chain 1	Group	9	0.3	Cyt, mem	HPA021649
DNAI2	Dynein, axonemal, intermediate chain 2	Group	5	0.2	Mem	CAB006245
SNTN	Sentan, cilia apical structure protein	Group	19	5	Mem	HPA043322
TEKT1	Tektin 1	Group	8	1	Mem	HPA044444
RSPH4A	Radial spoke head 4 homolog A	Enh	6	2	Cyt, mem	HPA031196
FOXJ1	Forkhead box J1	Enh	8	1	Nucl	HPA005714
Proteins specific for mucus-secreting cells						
SLPI	Secretory leukocyte peptidase inhibitor	Group	711	0.2	Sec	HPA027774
SCGB1A1	Secretoglobin, family 1A, member 1	Mod	1804	39	Sec	HPA031828
BPIFA1	BPI fold containing family A, member 1	Group	2	1	Sec	CAB025669
ANKUB1	Ankyrin repeat and ubiquitin domain containing 1	Group	2	1	Sec	HPA053749
DCDC2B	Doublecortin domain containing 2B	Enh	3	2	Sec	HPA045832
Proteins specific for endothelial cells						
ACE	Angiotensin I converting enzyme 1	Group	71	0.3	Mem, sec	CAB002426
PRX	Periaxin	Enh	23	4	Mem	HPA001868
Proteins specific for immune cells						
MSR1	Macrophage scavenger receptor 1	Enh	83	3	Cyt, mem	HPA000272
MARCO	Macrophage receptor with collagenous structure	Enh	131	2	Mem	HPA008847
MRC1	Mannose receptor, C type 1	Enh	50	2	Cyt, mem	HPA004114

(continued on next page)

TABLE 2. *(continued)*

Gene	Description	RNA category	mRNA level (FPKM)	Lung-specific score	Subcellular localization	HPA Ab example
C19ORF59	Chromosome 19 open reading frame 59	Group	64	2	Mem	HPA014731
TPSD1	Tryptase delta 1	Enh	8	1	Cyt	CAB002215
PCSK9	Proprotein convertase subtilisin/kexin type 9	Enh	9	1	Cyt	CAB025575
COL6A6	Collagen, type VI, α5	Enh	11	3	Cyt	HPA045239

RNA category indicates category of lung elevated expression, defined as highly lung enriched (high), moderately lung enriched (mod), group enriched (group), or lung enhanced (enh). Subcellular localization indicates localization of the immunohistochemical staining, defined as cytoplasmic (cyt), secreted (sec), membranous (mem), or nuclear (nucl). HPA Ab example indicates Human Protein Atlas ID of antibodies used in the immunohistochemically stained examples in Figs. 2–4.

domains, and overexpression has been shown to dramatically reduce the expression of CAV1 (29). In the present investigation, EMP2 was distinctly expressed in type I pneumocytes.

Analysis of proteins expressed in ciliated cells

Ciliated cells in the lung are present along bronchi and are important in maintaining the surface mucous membrane and freeing the airways from inhaled contaminants. In **Fig. 3**, examples of 16 proteins with expression in ciliated cells are shown. All these proteins were also expressed in the ciliated cells of other organs, such as the fallopian tube. Thus, only 2 genes (*CCDC17* and *FAM92B*) belong to the moderately enriched group, and the remaining 14 genes are group enriched or lung enhanced. For 9 genes (*C1ORF87*, *C9ORF171*, *CCDC17*, *CCDC19*, *CCDC37*, *CCDC42B*, *DYDC2*, *FAM92B*, and *FAM183A*), the function of the corresponding protein is unknown. Three genes (*DNAH5*, *DNAI1*, and *DNAI2*) encode the dynein proteins necessary for cilia movement and where defects cause diseases with severe pulmonary dysfunction, such as primary ciliary dyskinesia and Kartagener's syndrome (30). Three additional proteins (SNTN, TEKT1, and RSPH4A) showed distinct expression in cilia, out of which SNTN and TEKT1 had hitherto evidence of existence only at the transcript level. SNTN is a putative component of the linker structure that bridges ciliary membranes and peripheral singlet microtubules, and TEKT1 is a member of the tektin family that coassembles with tubulins to form ciliary and flagellar microtubules. RSPH4A is a probable component of the axonemal radial spoke head, regularly spaced along cilia, sperm, and flagella axonemes. Similar to the dynein proteins, dysfunction of this protein can cause primary ciliary dyskinesia (31). The transcription factor FOXJ1, the only example showing nuclear expression in ciliated cells, has been suggested to play a role in lung development and to regulate the transcription of genes controlling production of motile cilia.

Analysis of mucus-secreting cells in the lung

Mucus-secreting cells are present in many different tissues, and in lung, these cells reside in both the

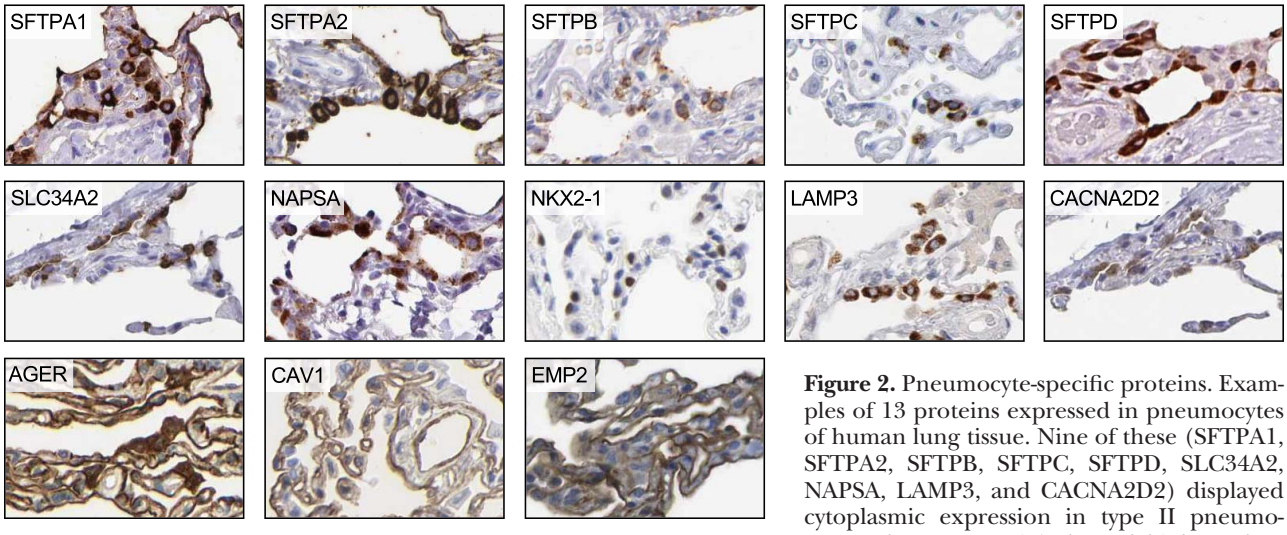

Figure 2. Pneumocyte-specific proteins. Examples of 13 proteins expressed in pneumocytes of human lung tissue. Nine of these (SFTPA1, SFTPA2, SFTPB, SFTPC, SFTPD, SLC34A2, NAPSA, LAMP3, and CACNA2D2) displayed cytoplasmic expression in type II pneumocytes, whereas NKX2-1 showed high nuclear expression in type II pneumocytes. AGER was distinctly expressed in type I pneumocytes, with additional weak positivity in endothelial cells. Expression of CAV1 and EMP2 was observed in plasma membranes of type I pneumocytes, accompanied with distinct expression in endothelial cells.

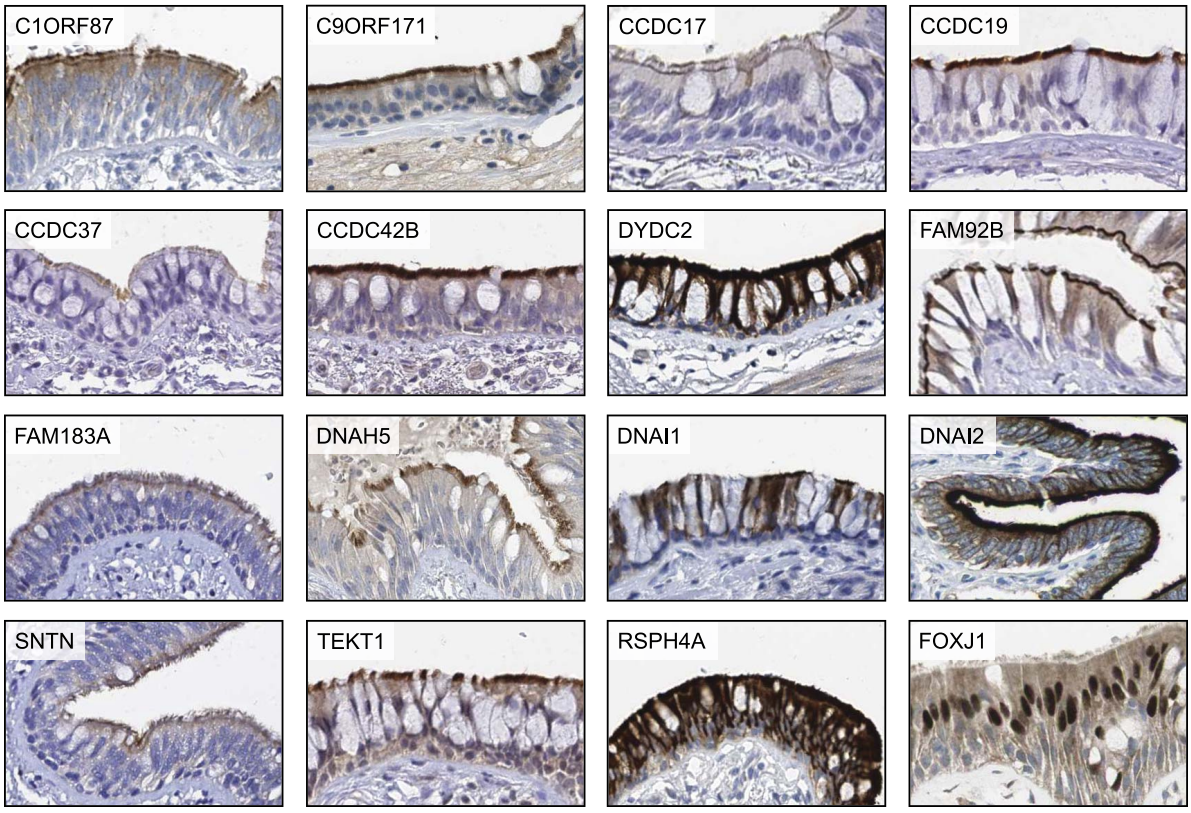

Figure 3. Proteins expressed in ciliated cells. Examples of 16 proteins expressed in ciliated cells in bronchial epithelium of human lung tissue. Ten of these proteins (C1ORF87, C9ORF171, CCDC17, CCDC19, CCDC37, CCDC42B, DNAH5, DNAI2, SNTN, and TEKT1) showed clear expression restricted to the cilia, of which 3 (C9ORF171, CCDC17, and SNTN) were selectively expressed at the most distal tip of the cilia. Three proteins (DNAI1, DYDC2, and RSPH4) displayed cytoplasmic expression of ciliated cells, accompanied with positivity of the cilia. FAM92B and FAM183A were expressed in the membrane under the cilia, whereas FOXJ1 showed nuclear expression of ciliated cells.

surface respiratory epithelium and the peribronchial glands. Secreted mucus is important for maintaining a suitable environment for ciliary function and protection against airborne infectious agents and solid particles. In **Fig. 4*A***, 5 examples of proteins specific for mucus-secreting cells in airways are shown. SLPI is a secreted inhibitor with broad-spectrum antibiotic activity, protecting epithelial tissues from serine proteases and attack by endogenous enzymes. SCGB1A1 is a member of the secretoglobin family, implicated in numerous functions, including anti-inflammation, and critical for epithelial regeneration after oxidant-induced injury. Defects in SCGB1A1 are associated with a susceptibility to asthma. Three less well known proteins distinctly expressed in airway mucus-secreting cells include BPIFA1, ANKUB1, and DCDC2B, the latter two with unknown function and with existence only at the transcript level, according to present evidence. BPIFA1 may be involved in inflammatory responses to irritants in the upper airways (32), with expression only in a smaller subset of mucus-producing cells in peribronchial glands and nasopharynx. ANKUB1 displayed additional expression in endothelial cells of several tissue types, including lung, whereas DCDC2B also was positive in mucus-secreting cells of gastric mucosa.

Endothelial cells in the lung

Vascular tissue and endothelial cells are present throughout the human body, with different structure and requirements depending on the specific function of respective tissue. In the lung, up to 30% of the cells represent specialized vasculature (33), lining the alveoli and participating in gaseous exchange. Two examples of proteins expressed in lung endothelial cells are displayed in Fig. 4*B*. ACE plays a key role in the renin–angiotensin system, involved in conversion of angiotensin I into physiologically active angiotensin II, a potent vasopressor and aldosterone-stimulating peptide controlling blood pressure and fluid–electrolyte balance. Expression of ACE was found in endothelial cells of both alveolar capillaries and larger arteries in lung tissue, as well as in blood vessels of other organs, except the renal glomeruli. PRX encodes a protein suggested to be required for maintenance of peripheral nerve myelin sheath, also playing a role in axon–glial interaction (34). In addition to expression in Schwann cells, the expression pattern was similar to that of ACE in lung; however, in other organs, expression was observed only in the renal glomeruli.

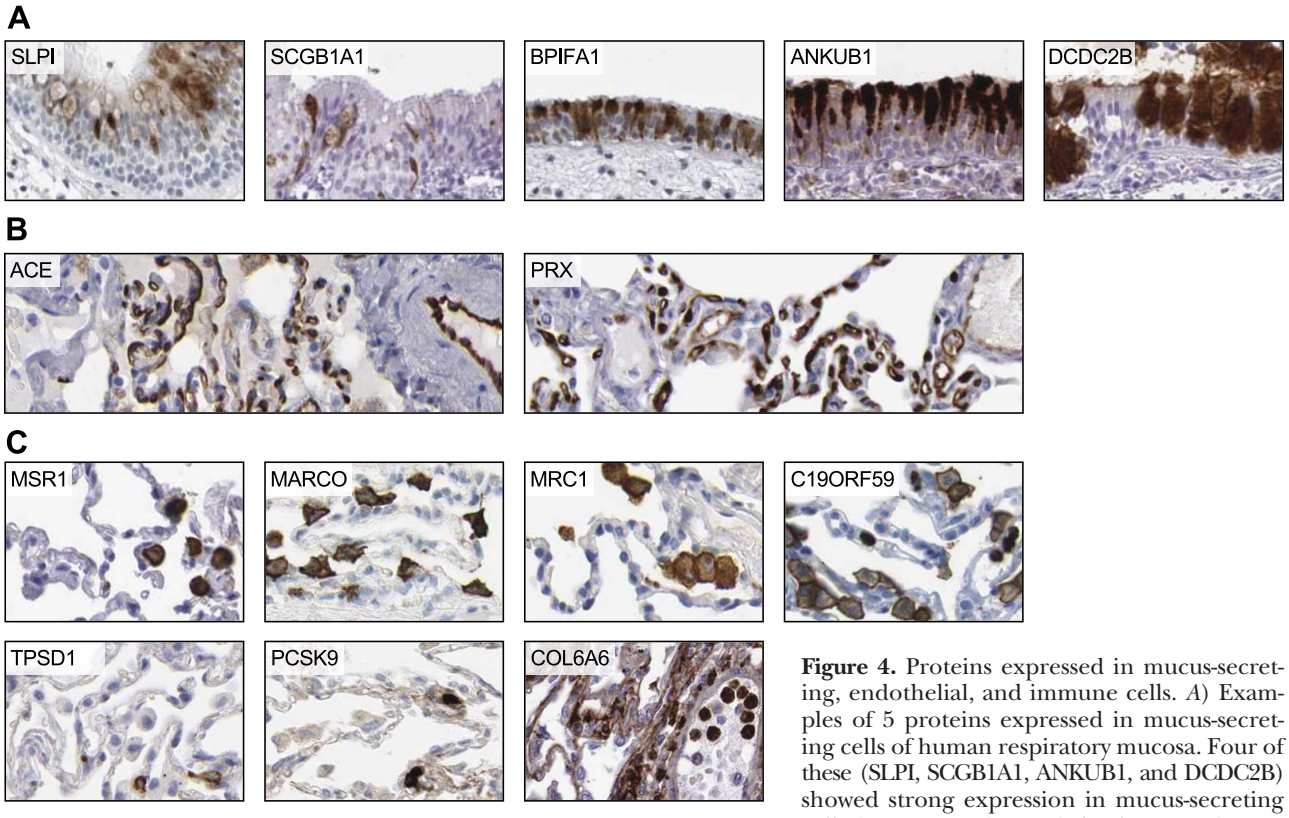

Figure 4. Proteins expressed in mucus-secreting, endothelial, and immune cells. *A*) Examples of 5 proteins expressed in mucus-secreting cells of human respiratory mucosa. Four of these (SLPI, SCGB1A1, ANKUB1, and DCDC2B) showed strong expression in mucus-secreting cells in most regions of the human airways, whereas BPIFA1was expressed only in mucus-secreting cells of the nasopharynx. *B*) Examples of 2 proteins expressed in endothelial cells of human lung. ACE and PRX showed selective membranous expression in lung capillaries and arteries, whereas veins were negative. *C*) Examples of 7 proteins expressed in immune cells of human lung. Four of these (MSR1, MARCO, MRC1, and C19ORF59) showed distinct membranous expression in alveolar macrophages, whereas TPSD1 and PCSK9 were expressed in the cytoplasm of mast cells. COL6A6 displayed clear expression of granulocytes, with additional expression observed in pneumocytes.

Immune cells in the lung

Seven genes identified as group enriched or enhanced in the lung tissue encoded proteins expressed in immune cells present in lung (Fig. 4*C*). Four of these (MSR1, MARCO, MRC1, and C19ORF59) were mainly expressed in macrophages, and 3 proteins (TPSD1, PCSK9, and COL6A6) were primarily expressed in other immune cell types. Examples of macrophage-specific expression include MSR1, a macrophage scavenger receptor that mediates endocytosis of macromolecules and is implicated in several macrophage-associated physiological and pathological processes, including atherosclerosis and host defense; MARCO, a scavenger receptor part of the innate antimicrobial immune system, that may bind both gram-negative and gram-positive bacteria; MRC1, a type I mannose receptor mediating endocytosis of glycoproteins by macrophages, binding to surface structures on pathogenic viruses, bacteria, and fungi to facilitate phagocytic engulfment; and C19ORF59, a protein with unknown function suggested to be expressed in mast cells, also showing membranous expression in alveolar macrophages. MARCO and MSR1 also displayed positivity in macrophages of other organs, such as the spleen, while MRC1 and C19ORF59 were demonstrated to be exclusively expressed in alveolar macrophages. Three proteins expressed in other inflammatory cell types include TPSD1, a tryptase secreted during the activation–degranulation response of mast cells; PCSK9, a secreted peptidase implicated in the regulation of plasma cholesterol homeostasis, also displaying distinct expression in mast cells; and COL6A6, a cell-binding collagen protein expressed in the extracellular matrix. In the present investigation, COL6A6 showed distinct expression in alveolar walls and granulocytes. Strong immunoreactivity was also displayed in myelopoietic cells of bone marrow and granulocytes present in other tissue types. TPSD1 and PCSK9 were expressed in mast cells of several different tissue types in addition to expression in lung. However, considering the relatively large number of mast cells in lung playing an important role in both health and disease, it is not surprising that mast-cell-specific expression showed a significant contribution to the RNA pool derived from lung tissue.

Prognostic relevance of the identified pneumocyte-specific genes in lung cancer

Based on the assumption that organ-specific gene expression reflects tissue homeostasis under physiological conditions, deregulation may accordingly be indicative of disease. To investigate how altered pneumocyte-specific expression patterns are related to lung cancer

prognosis, we used publically available gene expression microarray data from 7 independent patient cohorts, including altogether 1252 patients with NSCLC for whom there was accompanying information on overall survival. Of the 13 pneumocyte-specific genes, 11 were represented by ≥1 probe set on the Affymetrix U133A, U133B, or U133 Plus 2.0 microarray.

Using a meta-analysis approach to combine the results from the different cohorts into 1 pooled estimate, lower expression values based on probe sets of 7 pneumocyte-specific genes (*SFTPB*, *SFTPC*, *SFTPD*, *SLC34A2*, *LAMP3*, *CACNA2D2*, and *AGER*) was found to be significantly associated with shorter overall survival (adjusted $P<0.05$) when all histological NSCLC subtypes were included in the combined analysis (Supplemental Table S3). The survival association was even more pronounced when the adenocarcinoma subgroup was analyzed separately (**Table 3**). In addition, probe sets of *NAPSA*, *NKX2-1*, and *EMP2* were associated with overall survival in this histological subtype. Representative results for the adenocarcinoma patient subset are illustrated with forest plots (**Fig. 5**). No prognostic effect of pneumocyte-specific gene expression was observed for squamous cell lung carcinoma (Supplemental Table S3).

TABLE 3. *Eleven pneumocyte-specific genes (25 probe sets) and their association with prognosis in lung adenocarcinoma based on a meta-analysis of gene expression microarray data from 6 independent patient cohorts*

					95% CI	
Gene	Affymetrix ID	HR	*P*	*P*, FDR	Lower	Upper
SFTPB	209810_at	0.90	0.0000	0.0003	0.86	0.95
	213936_x_at	0.87	0.0000	0.0000	0.82	0.92
	214354_x_at	0.88	0.0000	0.0000	0.83	0.93
	37004_at	0.91	0.0001	0.0005	0.87	0.95
SFTPC	205982_x_at	0.94	0.0008	0.0019	0.91	0.97
	211735_x_at	0.94	0.0006	0.0015	0.90	0.97
	214387_x_at	0.93	0.0004	0.0010	0.90	0.97
	215454_x_at	0.87	0.0002	0.0006	0.81	0.94
	38691_s_at	0.94	0.0013	0.0025	0.90	0.98
SFTPD	214199_at	0.90	0.0002	0.0006	0.85	0.95
SLC34A2	204124_at	0.87	0.0003	0.0008	0.80	0.94
NAPSA	223806_s_at	0.89	0.0044	0.0078	0.83	0.97
NKX2-1	210673_x_at	0.82	0.0695	0.0869	0.66	1.02
	211024_s_at	0.87	0.0012	0.0025	0.80	0.95
	231315_at	0.85	0.0144	0.0212	0.75	0.97
LAMP3	205569_at	0.90	0.0077	0.0128	0.84	0.97
CACNA2D2	204811_s_at	0.86	0.0000	0.0001	0.80	0.92
AGER	210081_at	0.91	0.0363	0.0478	0.83	0.99
	217046_s_at	0.84	0.0125	0.0195	0.73	0.96
CAV1	203065_s_at	0.94	0.2038	0.2215	0.86	1.03
	212097_at	0.96	0.5392	0.5616	0.86	1.08
EMP2	204975_at	0.87	0.0228	0.0317	0.76	0.98
	225078_at	0.86	0.1015	0.1209	0.72	1.03
	225079_at	0.84	0.1135	0.1289	0.68	1.04
	238500_at	0.89	0.6322	0.6322	0.55	1.44

HR, hazard ratio; FDR, false discovery rate; CI, confidence interval.

DISCUSSION

Our approach combines RNA-Seq data describing the lung-specific transcriptome with cell-type-specific evaluation of corresponding protein expression in normal human lung. The analysis identified 20 genes that are clearly enriched in the lung and 201 genes with an enhanced expression or a shared expression with ≥1 other tissue type.

Previously, several attempts have been made to describe the organ-specific molecular repertoire with gene expression profiles reflecting average mRNA levels over all cell types (5, 35, 36). The presence of immune cells, fibroblasts, and cells of the vasculature naturally dilutes the expression profiles of epithelial parenchyma. In lung, these cells represent an abundant and functionally important tissue element. To address this intrinsic cellular heterogeneity, the mRNA levels were compared to corresponding *in situ* protein expression by visual examination of immunohistochemically stained images, allowing further stratification according to cell type and subcellular compartment. In particular, we focused on proteins expressed specifically in pneumocytes and in ciliated, mucus-secreting, endothelial, and immune cells.

The list of lung-specific mRNA transcripts comprised, not unexpectedly, several well-known genes of particular importance in lung homeostasis and forming of the surface film essential for normal respiration, such as the surfactants A, B, C, and D expressed in type II pneumocytes. Two genes, *NKX2-1* and *NAPSA*, are already included in diagnostic pathology to evaluate the origin of adenocarcinoma in the lung and metastases (23). Notably, all proteins selectively expressed in bronchial epithelium showed low or absent mRNA expression in the sample that microscopically lacked bronchial epithelium, indicating high sensitivity of the technique. In addition to the well-known lung markers, several proteins hitherto not characterized in lung were identified. One example is EMP2, a membrane protein thought to interact with CAV1 (29), which is known to be expressed in type I pneumocytes; however, the exact function of EMP2 is unknown. Another example is PRX, associated with sciatic nerves and Schwann cells (34), but also displaying distinct expression in a subset of lung endothelial cells that has never been described. Moreover, several proteins with completely unknown function were identified, including several proteins expressed in ciliated cells, such as CCDC17 and FAM92B, and the putative mast cell protein C19ORF59, showing specific membranous expression in alveolar macrophages in the present investigation. It is beyond the scope of our study to determine the exact function of each identified protein, but an important implication in lung homeostasis can be anticipated and should be further investigated.

The importance of our study goes beyond the provision of mere descriptive information, as the data can serve as a starting point for biomarker discovery. Novel biomarkers may have a significant clinical effect in the

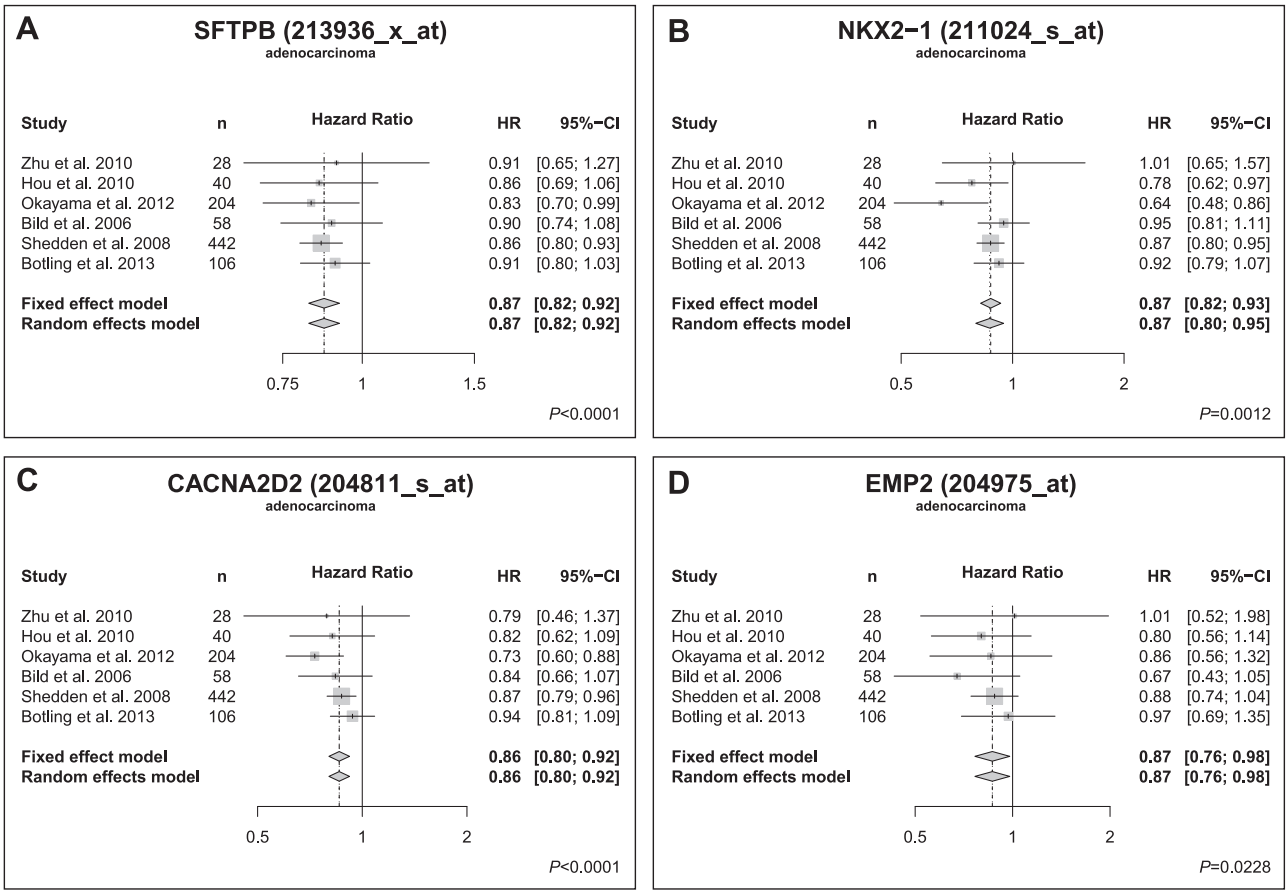

Figure 5. Prognostic relevance of pneumocyte-specific genes in lung adenocarcinoma. Forest plots illustrating the results of the meta-analysis for SFTPB (*A*), NKX2-1 (*B*), CACNA2D2 (*C*), and EMP2 (*D*) in lung adenocarcinoma. Studies in meta-analysis: Zhu *et al.* (16), Okyama *et al.* (17), Bild *et al.* (18), Sheddon *et al.* (2), and Botling *et al.* (3).

context of disease screening and risk stratification, as well as prediction of therapy response and side effects. Postulating that tissue-specific gene expression is indicative of normal organ function and that deregulation consequently may be suggestive of a disease state, we explored this concept by investigating the expression of pneumocyte-specific genes in relation to lung cancer prognosis. We found that higher expression of probe sets of 7 genes (*SFTPB*, *SFTPC*, *SFTPD*, *SLC34A2*, *LAMP3*, *CACNA2D2*, and *AGER*) was significantly associated with longer overall survival when all histological subtypes were included in the statistical model, and significance was maintained after adjustment for multiple testing. The survival association was even more pronounced when patients with adenocarcinoma were evaluated separately, and *NAPSA*, *NKX2-1*, and *NAPSA* also revealed a significant prognostic effect in this histological subtype. However, prognostic value does not necessarily imply direct involvement in lung cancer tumorigenesis. It can be speculated that the presented biomarker candidates are only surrogates for the degree of differentiation and thus reflects the aggressiveness of the tumor. That deregulation of pneumocyte-specific markers shows only a prognostic effect in adenocarcinoma supports the hypothesis that squamous cell cancer has another cell of origin (*i.e.*, not only pneumocytes, but the bronchial epithelial cells after metaplasia; refs. 37–39).

Only two of the aforementioned genes have been associated with survival outcome, including the lung adenocarcinoma markers, well-known to pathologists, *NKX2-1* (alias *TTF-1*; ref. 40) and *NAPSA* (*Napsin-A*; ref. 41), whereas the other 8 genes have not been described in this context. Surfactant proteins B and C (*SFTPB*, *SFTPC*) have been reported as biomarkers of the presence of lung cancer in serum and in lymph node metastases (42, 43), and a genetic variant of *SFTPD* has been linked to lung cancer risk in a Japanese cohort (44). However, none of the surfactant proteins has been linked to lung cancer survival. Sodium phosphate cotransporter gene type IIb (*SLC34A2*) has been become a recent focus of clinical oncology as a fusion partner for the tyrosine kinase *ROS1*, which is now the object of targeted treatment (45), but a prognostic effect has not been reported until today. The α-2/δ subunit of voltage-dependent calcium channel (*CACNA2D2*) has been shown to be frequently deleted in lung cancer, and tumor suppressor functions have been proposed (46). In the current study, we have shown as a clinical correlate that higher expression of *CACNA2D2* is indeed associated with longer survival. Also, the 3 other identified pneumocyte-specific proteins, AGER, LAMP3, and EMP2, all

with prognostic effect in our analysis, have not been described previously as biomarkers in lung cancer.

To our knowledge, this study is the first to use the RNA-Seq technique to describe normal human lung and identify lung-specific gene expression and corresponding protein expression. It should be noted that some aspects of our strategy have limitations. The number of normal lung cases (n=5) is a relatively small sample, and all tissues were obtained after lung surgery, suggesting that underlying diseases may influence the results. Furthermore, we were not able to confirm the expression on the protein level and to assign the expression to a certain cell type for all of the genes identified as elevated in lung. The reasons for this are various: gene expression does translate into protein expression; the protein amount is still under the immunohistochemical detection limit; or the antibodies used do not provide sufficient sensitivity or specificity for protein detection. However, for the purpose of localizing protein expression in the *in situ* environment of the lung, we regard the immunohistochemical approach as the best alternative. With the advancement of the Human Protein Atlas and the generation of new well validated antibodies, the genes that can be confirmed on the protein level will increase.

In summary, our study provides a detailed transcriptomic overview, characterizing normal lung tissue in relationship to a broad range of other organs and tissues. The generated data set offers a unique resource for hypothesis generation and confirmation and gives access to antibody tools for further lung-related studies. Furthermore, the presented findings demonstrate a potential clinical relevance of pneumocyte-specific gene expression. FJ

The authors thank the entire staff of the Human Protein Atlas program and the Science for Life Laboratory for valuable contributions, and the Department of Pathology at the Uppsala Akademiska Hospital and the Uppsala Biobank for kindly providing the clinical diagnostics and specimens used in this study. Funding was provided by the Knut and Alice Wallenberg Foundation and PROSPECTS, a 7th Framework grant by the European Directorate (grant agreement HEALTH-F4-2008–201648/PROSPECTS), and the Swedish Cancerfonden. Parts of this work were also supported by research grants from the Lions Cancerforskningsfond and the Swedish Cancer Society (both to P.M). The work of B.H. was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, grant RA 870/5-1).

REFERENCES

1. Campbell, J. D., Spira, A., and Lenburg, M. E. (2011) Applying gene expression microarrays to pulmonary disease. *Respirology* **16,** 407–418
2. Shedden, K., Taylor, J. M., Enkemann, S. A., Tsao, M. S., Yeatman, T. J., Gerald, W. L., Eschrich, S., Jurisica, I., Giordano, T. J., Misek, D. E., Chang, A. C., Zhu, C. Q., Strumpf, D., Hanash, S., Shepherd, F. A., Ding, K., Seymour, L., Naoki, K., Pennell, N., Weir, B., Verhaak, R., Ladd-Acosta, C., Golub, T., Gruidl, M., Sharma, A., Szoke, J., Zakowski, M., Rusch, V., Kris, M., Viale, A., Motoi, N., Travis, W., Conley, B., Seshan, V. E., Meyerson, M., Kuick, R., Dobbin, K. K., Lively, T., Jacobson, J. W., and Beer, D. G.; Director's Challenge Consortium for the Molecular Classification of Lung (2008) Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat. Med.* **14,** 822–827
3. Botling, J., Edlund, K., Lohr, M., Hellwig, B., Holmberg, L., Lambe, M., Berglund, A., Ekman, S., Bergqvist, M., Ponten, F., Konig, A., Fernandes, O., Karlsson, M., Helenius, G., Karlsson, C., Rahnenfuhrer, J., Hengstler, J. G., and Micke, P. (2013) Biomarker discovery in non-small cell lung cancer: integrating gene expression profiling, meta-analysis, and tissue microarray validation. *Clin. Cancer Res.* **19,** 194–204
4. Boutros, P. C., Lau, S. K., Pintilie, M., Liu, N., Shepherd, F. A., Der, S. D., Tsao, M. S., Penn, L. Z., and Jurisica, I. (2009) Prognostic gene signatures for non-small-cell lung cancer. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 2824–2828
5. Wang, Z., Gerstein, M., and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63
6. Fagerberg, L., Hallstrom, B. M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., Habuka, M., Tahmasebpoor, S., Danielsson, A., Edlund, K., Asplund, A., Sjostedt, E., Lundberg, E., Szigyarto, C. A., Skogs, M., Takanen, J. O., Berling, H., Tegel, H., Mulder, J., Nilsson, P., Schwenk, J. M., Lindskog, C., Danielsson, F., Mardinoglu, A., Sivertsson, A., von Feilitzen, K., Forsberg, M., Zwahlen, M., Olsson, I., Navani, S., Huss, M., Nielsen, J., Ponten, F., and Uhlen, M. (2014) Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteomics* **13,** 397–406
7. Joshi, N. A., and Fass, J. N. (2012) Sickle: a windowed adaptive trimming tool for FASTQ files using quality. Available at https://github.com/najoshi/sickle
8. Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–515
9. Anonymous (2012) Picard sequence alignment tools. Available at http://picard.sourceforge.net/
10. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–2504
11. Hebenstreit, D., Fang, M., Gu, M., Charoensawan, V., van Oudenaarden, A., and Teichmann, S. A. (2011) RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Mol. Syst. Biol.* **7,** 497
12. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25,** 25–29
13. Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformat.* **10,** 48
14. Kampf, C., Olsson, I., Ryberg, U., Sjostedt, E., and Ponten, F. (2012) Production of tissue microarrays, immunohistochemistry staining and digitalization within the human protein atlas. *J. Vis. Exp.* **63,** pii 3620
15. Zhu, C. Q., Ding, K., Strumpf, D., Weir, B. A., Meyerson, M., Pennell, N., Thomas, R. K., Naoki, K., Ladd-Acosta, C., Liu, N., Pintilie, M., Der, S., Seymour, L., Jurisica, I., Shepherd, F. A., and Tsao, M. S. (2010) Prognostic and predictive gene signature for adjuvant chemotherapy in resected non-small-cell lung cancer. *J. Clin. Oncol.* **28,** 4417–4424
16. Hou, J., Aerts, J., den Hamer, B., van Ijcken, W., den Bakker, M., Riegman, P., van der Leest, C., van der Spek, P., Foekens, J. A., Hoogsteden, H. C., Grosveld, F., and Philipsen, S. (2010) Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PLoS One* **5,** e10312
17. Okayama, H., Kohno, T., Ishii, Y., Shimada, Y., Shiraishi, K., Iwakawa, R., Furuta, K., Tsuta, K., Shibata, T., Yamamoto, S., Watanabe, S., Sakamoto, H., Kumamoto, K., Takenoshita, S., Gotoh, N., Mizuno, H., Sarai, A., Kawano, S., Yamaguchi, R., Miyano, S., and Yokota, J. (2012) Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. *Cancer Res.* **72,** 100–111

18. Bild, A. H., Yao, G., Chang, J. T., Wang, Q., Potti, A., Chasse, D., Joshi, M. B., Harpole, D., Lancaster, J. M., Berchuck, A., Olson, J. A., Jr., Marks, J. R., Dressman, H. K., West, M., and Nevins, J. R. (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439,** 353–357
19. Raponi, M., Zhang, Y., Yu, J., Chen, G., Lee, G., Taylor, J. M., Macdonald, J., Thomas, D., Moskaluk, C., Wang, Y., and Beer, D. G. (2006) Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. *Cancer Res.* **66,** 7466–7472
20. Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4,** 249–264
21. Benjamini, Y., and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Methodol.* **57,** 289–300
22. Corut, A., Senyigit, A., Ugur, S. A., Altin, S., Ozcelik, U., Calisir, H., Yildirim, Z., Gocmen, A., and Tolun, A. (2006) Mutations in SLC34A2 cause pulmonary alveolar microlithiasis and are possibly associated with testicular microlithiasis. *Am. J. Hum. Genet.* **79,** 650–656
23. Ueno, T., Linder, S., and Elmberger, G. (2003) Aspartic proteinase napsin is a useful marker for diagnosis of primary lung adenocarcinoma. *Br. J. Cancer* **88,** 1229–1233
24. Mulugeta, S., Gray, J. M., Notarfrancesco, K. L., Gonzales, L. W., Koval, M., Feinstein, S. I., Ballard, P. L., Fisher, A. B., and Shuman, H. (2002) Identification of LBM180, a lamellar body limiting membrane protein of alveolar type II cells, as the ABC transporter protein ABCA3. *J. Biol. Chem.* **277,** 22147–22155
25. Gong, H. C., Hang, J., Kohler, W., Li, L., and Su, T. Z. (2001) Tissue-specific expression and gabapentin-binding properties of calcium channel alpha2delta subunit subtypes. *J. Membr. Biol.* **184,** 35–43
26. Carboni, G. L., Gao, B., Nishizaki, M., Xu, K., Minna, J. D., Roth, J. A., and Ji, L. (2003) CACNA2D2-mediated apoptosis in NSCLC cells is associated with alterations of the intracellular calcium signaling and disruption of mitochondria membrane integrity. *Oncogene* **22,** 615–626
27. Hofmann, M. A., Drury, S., Fu, C., Qu, W., Taguchi, A., Lu, Y., Avila, C., Kambham, N., Bierhaus, A., Nawroth, P., Neurath, M. F., Slattery, T., Beach, D., McClary, J., Nagashima, M., Morser, J., Stern, D., and Schmidt, A. M. (1999) RAGE mediates a novel proinflammatory axis: a central cell surface receptor for S100/calgranulin polypeptides. *Cell* **97,** 889–901
28. Ho, C. C., Kuo, S. H., Huang, P. H., Huang, H. Y., Yang, C. H., and Yang, P. C. (2008) Caveolin-1 expression is significantly associated with drug resistance and poor prognosis in advanced non-small cell lung cancer patients treated with gemcitabine-based chemotherapy. *Lung Cancer* **59,** 105–110
29. Wadehra, M., Goodglick, L., and Braun, J. (2004) The tetraspan protein EMP2 modulates the surface expression of caveolins and glycosylphosphatidyl inositol-linked proteins. *Mol. Biol. Cell* **15,** 2073–2083
30. Zariwala, M. A., Knowles, M. R., and Omran, H. (2007) Genetic defects in ciliary structure and function. *Annu. Rev. Physiol.* **69,** 423–450
31. Castleman, V. H., Romio, L., Chodhari, R., Hirst, R. A., de Castro, S. C., Parker, K. A., Ybot-Gonzalez, P., Emes, R. D., Wilson, S. W., Wallis, C., Johnson, C. A., Herrera, R. J., Rutman, A., Dixon, M., Shoemark, A., Bush, A., Hogg, C., Gardiner, R. M., Reish, O., Greene, N. D., O'Callaghan, C., Purton, S., Chung, E. M., and Mitchison, H. M. (2009) Mutations in radial spoke head protein genes RSPH9 and RSPH4A cause primary ciliary dyskinesia with central-microtubular-pair abnormalities. *Am. J. Hum. Genet.* **84,** 197–209
32. Liu, Y., Bartlett, J. A., Di, M. E., Bomberger, J. M., Chan, Y. R., Gakhar, L., Mallampalli, R. K., McCray, P. B., Jr., and Di, Y. P. (2013) SPLUNC1/BPIFA1 contributes to pulmonary host defense against Klebsiella pneumoniae respiratory infection. *Am. J. Pathol.* **182,** 1519–1531
33. Crapo, J. D., Barry, B. E., Gehr, P., Bachofen, M., and Weibel, E. R. (1982) Cell number and cell characteristics of the normal human lung. *Am. Rev. Respir. Dis.* **126,** 332–337
34. Gillespie, C. S., Sherman, D. L., Blair, G. E., and Brophy, P. J. (1994) Periaxin, a novel protein of myelinating Schwann cells with a possible role in axonal ensheathment. *Neuron* **12,** 497–508
35. Jongeneel, C. V., Delorenzi, M., Iseli, C., Zhou, D., Haudenschild, C. D., Khrebtukova, I., Kuznetsov, D., Stevenson, B. J., Strausberg, R. L., Simpson, A. J., and Vasicek, T. J. (2005) An atlas of human gene expression from massively parallel signature sequencing (MPSS). *Genome Res.* **15,** 1007–1014
36. Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R., and Hogenesch, J. B. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 6062–6067
37. Lantuejoul, S., Salameire, D., Salon, C., and Brambilla, E. (2009) Pulmonary preneoplasia: sequential molecular carcinogenetic events. *Histopathology* **54,** 43–54
38. Sutherland, K. D., and Berns, A. (2010) Cell of origin of lung cancer. *Mol. Oncol.* **4,** 397–403
39. Wistuba, I. I., Behrens, C., Milchgrub, S., Bryant, D., Hung, J., Minna, J. D., and Gazdar, A. F. (1999) Sequential molecular abnormalities are involved in the multistage development of squamous cell lung carcinoma. *Oncogene* **18,** 643–650
40. Solis, L. M., Behrens, C., Raso, M. G., Lin, H. Y., Kadara, H., Yuan, P., Galindo, H., Tang, X., Lee, J. J., Kalhor, N., Wistuba, I. I., and Moran, C. A. (2012) Histologic patterns and molecular characteristics of lung adenocarcinoma associated with clinical outcome. *Cancer* **118,** 2889–2899
41. Lee, J. G., Kim, S., and Shim, H. S. (2012) Napsin A is an independent prognostic factor in surgically resected adenocarcinoma of the lung. *Lung Cancer* **77,** 156–161
42. Sin, D. D., Tammemagi, C. M., Lam, S., Barnett, M. J., Duan, X., Tam, A., Auman, H., Feng, Z., Goodman, G. E., Hanash, S., and Taguchi, A. (2013) Pro-surfactant protein B as a biomarker for lung cancer prediction. *J. Clin. Oncol.* **31,** 4536–4543
43. Nordgard, O., Singh, G., Solberg, S., Jorgensen, L., Halvorsen, A. R., Smaaland, R., Brustugun, O. T., and Helland, A. (2013) Novel molecular tumor cell markers in regional lymph nodes and blood samples from patients undergoing surgery for non-small cell lung cancer. *PLoS One* **8,** e62153
44. Ishii, T., Hagiwara, K., Ikeda, S., Arai, T., Mieno, M. N., Kumasaka, T., Muramatsu, M., Sawabe, M., Gemma, A., and Kida, K. (2012) Association between genetic variations in surfactant protein d and emphysema, interstitial pneumonia, and lung cancer in a Japanese population. *COPD* **9,** 409–416
45. Rimkunas, V. M., Crosby, K. E., Li, D., Hu, Y., Kelly, M. E., Gu, T. L., Mack, J. S., Silver, M. R., Zhou, X., and Haack, H. (2012) Analysis of receptor tyrosine kinase ROS1-positive tumors in non-small cell lung cancer: identification of a FIG-ROS1 fusion. *Clin. Cancer Res.* **18,** 4449–4457
46. Hesson, L. B., Cooper, W. N., and Latif, F. (2007) Evaluation of the 3p21.3 tumour-suppressor gene cluster. *Oncogene* **26,** 7283–7301

Received for publication April 4, 2014.
Accepted for publication August 18, 2014.