Gene Expression Profiles of Non-Small Cell Lung Cancer: Survival Prediction and New Biomarkers

Kristjan Vääk^{a, d} Tõnu Vooder^{a, d, e} Raivo Kolde^b Mari-Ann Reintam^f
Cathleen Petzold^g Jaak Vilo^b Andres Metspalu^{a, c, d}

Institutes of ^aMolecular and Cell Biology and ^bComputer Science, ^cEstonian Genome Center, and ^dEstonian Biocentre, University of Tartu, and ^eClinic of Cardiovascular and Thoracic Surgery and ^fDepartment of Pathology, Tartu University Hospital, Tartu, Estonia; ^gCenter for Regenerative Therapies Dresden, Dresden, Germany

Key Words

Lung cancer, non-small cell • Survival prediction • Gene expression profile • RNA degradation

Abstract

Objectives: Despite the well-defined histological types of non-small cell lung cancer (NSCLC), a given stage is often associated with wide-ranging survival rates and treatment outcomes. This disparity has led to an increased demand for the discovery and identification of new informative biomarkers. **Methods:** In the current study, we screened 81 NSCLC samples using Illumina® whole-genome gene expression microarrays in an effort to identify differentially expressed genes and new NSCLC biomarkers. **Results:** We identified novel genes whose expression was upregulated in NSCLC, including SPAG5, POLH, KIF23, and RAD54L, which are associated with mitotic spindle formation, DNA repair, chromosome segregation, and dsDNA break repair, respectively. We also identified several novel genes whose expression was downregulated in NSCLC, including SGCG, NLRC4, MMRN1, and SFTPD, which are involved in extracellular matrix formation, apoptosis, blood vessel leakage, and inflammation, respectively. We found a significant correlation between RNA degradation and survival in adenocarcinoma cases. **Conclusions:** Even though the follow-up time was too limited to

draw final conclusions, we were able to show better prediction p values in a group selection based on molecular profiles compared to histology. The current study also uncovered new candidate biomarker genes that are likely to be involved in diverse processes associated with NSCLC development.

Copyright © 2011 S. Karger AG, Basel

Introduction

Lung cancer has the highest mortality rate among all types of cancers, resulting in more annual deaths than breast, prostate, colorectal, and pancreatic cancers combined [1].

Conventional diagnosis of lung cancer is currently based on tumor histology. The vast majority of lung cancers, approximately 80%, are non-small cell lung cancers (NSCLCs). The main histological types of NSCLC include adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. Despite the well-defined histological types of NSCLC, patient survival and treatment outcomes can dif-

K.V. and T.V. contributed equally to this work and share first authorship.

fer substantially, even among NSCLCs of the same histological type and stage [2]. Lung cancer morphology exhibits a broad spectrum and many tumors are atypical or lack the morphologic features necessary for improved differential diagnoses. Lung cancer diagnoses based solely on morphological features are in many cases insufficient [3].

In recent years, there has been an expanding interest in the molecular genetic classification of cancers. Gene expression profiling is one of the methods used for this purpose. The data obtained from gene expression analyses may allow the differentiation of cancer subgroups based on molecular phenotypes. Subgroups determined by gene expression in combination with clinical features like spread reflect the cancer's biology and progression better than histology alone [4].

Many genetic alterations at both the DNA and RNA levels have been shown to be associated with the development and progression of lung cancer. However, the precise underlying molecular mechanisms associated with NSCLC remain elusive in most cases. In the current study, we analyzed NSCLC specimens using Illumina® whole-genome gene expression microarrays and discovered novel biomarkers of NSCLC.

Patients and Methods

Patients and Samples

Initially, 146 patients diagnosed with NSCLC underwent surgery at the Clinic of Cardiovascular and Thoracic Surgery of Tartu University Hospital, Estonia. All patients gave their written informed consent to participate in the survey and to allow their biological samples to be genetically analyzed. The Ethics Review Committee on Human Research of the University of Tartu approved the current study. A patient questionnaire, as well as the approval to participate in the study and to provide control and tumor samples were obtained from all subjects. Tumor specimens and control samples were collected during surgery from November 28, 2002 to December 31, 2006, and the departmental pathologist promptly examined the specimens. Tumor histology and stage were estimated according to World Health Organization (WHO) [5] and TNM (tumour, node, and metastasis) staging according to UICC (International Union Against Cancer) classifications [6]. The same pathologist performed the histological classifications. Control samples were obtained from the same cancer patient at a site distant from the tumor and were approved as control samples by the pathologist. After the final diagnosis, a substantial number of patients were excluded from the gene expression study. The exclusion criteria and the number of patients excluded from the study are shown in table 1. From the initial cohort of 146 samples, patients who had received preoperative chemoradiotherapy and patients whose final histology did not show NSCLC were excluded.

After the exclusion of patients based on their medical records and final histology, 131 samples were available for the RNA extraction. When necessary, the RNA was extracted twice to meet the

Table 1. Patient exclusion criteria and the number of samples excluded

Exclusion criteria	Samples eliminated, n
Preoperative chemoradiotherapy	10
Other types of cancers	7
Nonmalignant lesions (fibrosis, tuberculosis, etc.)	9
RNA degradation	50

predefined RNA integrity number (RIN) cutoff value of ≥ 7 . Due to rapid RNA degradation among many of the cancer specimens, an additional 50 samples were excluded (the overall RNA recovery rate was 61.8%). After the RIN cutoff value was applied, a total of 81 samples were available for the gene expression experiments.

The age of the NSCLC patients enrolled in the gene expression study ranged from 38 to 81 years (mean 65.8). The gender distribution was 9 females and 72 males (11 and 89%, respectively). According to the histology, there were 13 bronchioloalveolar carcinomas (a subtype of adenocarcinoma), 8 adenocarcinomas, and 60 squamous cell carcinomas (16, 10, and 74%, respectively). Detailed characteristics of the patients and tumors analyzed are provided in table 2.

RNA Isolation and Amplification

Postsurgical tissue specimens were immediately cut to an appropriate size (maximum 1 cm³) and submerged in liquid nitrogen to inhibit RNA degradation. The samples were stored at -80°C until further processing. If necessary, the tissue samples were cut once more to meet the requirements for RNeasy®-Lysis (catalog No. AM7030; Ambion) treatment before RNA extraction.

Total cellular RNA from tissue specimens of 50 mg was extracted and purified using a Ribopure Kit (catalog No. AM1924; Ambion) according to the manufacturer's instructions. For tissue disruption, an IKA Ultra-Turrax T8 homogenizer was used. RNA quantity was assessed using a NanoDrop 1000 spectrophotometer, and Agilent Bioanalyzer lab-on-a-chip technology (Agilent RNA 6000 Nano Kit, catalog No. 5067-1511) was used to estimate RNA integrity. An RIN cutoff value of 7 was applied. An Illumina TotalPrep RNA Amplification Kit (catalog No. AMIL1971; Ambion) was used for RNA amplification and labeling. Amplifications were carried out according to the manufacturer's instructions using 300 ng of total RNA as a template.

Gene Expression Microarray

We used an Illumina BeadChip platform and corresponding Human-6 Expression Whole-Genome arrays containing more than 48,000 transcript probes for the microarray gene expression analysis. Experiments were carried out according to the manufacturer's instructions, and 1.5 µg of amplified cRNA was hybridized per single array. Slides were scanned immediately after the experiment using default settings, with the exception of the 'factor' setting which was set to 2.5. Illumina internal controls and BeadStudio software were used for data consistency and quality control of the hybridization data.

Statistical Analysis of Gene Expression Data

The expression data was quantile normalized and log transformed prior to the analysis to eliminate systematic differences between the chips, thereby standardizing the expression value distribution on all of the arrays. The aim of the analysis was to identify differentially expressed genes between NSCLC and control samples as well as within various types of cancerous tissues (e.g. lung adenocarcinoma vs. epidermoid cancer).

Differential gene expression analysis was performed using t tests with an empirical Bayes correction from the Bioconductor Limma package [7]. We used the Bonferroni correction for multiple testing and a significance level of $\alpha = 0.05$ in all comparisons. Gene Ontology (GO) enrichments were calculated using a g:Profiler web toolkit [8]. The statistically significant differentially expressed genes were clustered hierarchically using correlation distance and were visualized using a heatmap. In addition to the statistical parameters described previously, we used a minimum of a 2-fold change of expression to reduce the number of differentially expressed NSCLC-specific genes for further analysis.

Results

Gene Expression Profiles to Distinguish the Different Histological Subtypes of NSCLC

To distinguish between squamous cell lung cancer and adenocarcinoma and its subtype bronchioloalveolar carcinoma of the lung, gene selection was carried out using analysis of variance. Our overall goal was to identify genes that were differentially expressed between the 2 different types of lung cancer. The genes were selected using an F test, which shows how much of the expression is determined by a given histology. The F test p values were corrected for multiple testing using a false discovery rate and an empirical Bayes procedure provided by the Limma package of R software from Bioconductor. A significance level of 0.001 was applied, which resulted in the identification of 97 genes for further analysis.

Among the biomarkers that passed the filter, there were several recently described lung cancer-related genes, such as SERPINB3, CKS1B, PRAME, CDH2, CXCL6, TPX2, and MAGEA3 [9–15]. We also discovered a large number of novel NSCLC-associated biomarkers that had not been previously published as lung cancer associated, although some have reported roles in the development of other forms of cancer, including BUB1B, CRABP2, and TNFRSF18. To better visualize the results of this analysis, agglomerative hierarchical gene expression clustering was performed. Furthermore, the gene expression profiles of the control samples were included in the heat map (fig. 1).

We were able to distinguish most of the lung adenocarcinomas from the squamous cell lung cancers. Inter-

Table 2. Clinical and pathological characteristics of the patients enrolled in the gene expression study after the medical exclusions and RIN cutoff

	Patients for gene expression (n = 81)	%
Histology		
Adenocarcinoma	8	9.90
Bronchioloalveolar carcinoma	13	16.00
Squamous cell carcinoma	60	74.10
Lymph node		
Positive	13	16.00
Negative	68	84.00
Differentiation		
Well/moderately differentiated	76	94.00
Poorly/not differentiated	5	6.00
Stage		
Ia	13	16.10
Ib	46	56.80
IIa	1	1.20
IIb	3	3.70
IIIa	7	8.60
IIIb	6	7.40
IV	5	6.20
T1	15	18.50
T2	56	69.10
T3	5	6.20
T4	5	6.20
Tumor size		
<30, mm	36	44.40
>30, mm	45	55.60
Surgical procedure		
Wedge resection	6	7.40
Lobectomy	54	66.70
Bilobectomy	3	3.70
Pneumonectomy	18	22.20
Gender		
Female	9	11.10
Male	72	88.90
Age		
Range, years	38–81	
Mean, years	65.8	
Median, years	68	
<39 years	1	1.20
40–49 years	5	6.20
50–59 years	13	16.00
60–69 years	27	33.30
>70 years	35	43.20
Smoking		
Nonsmoker	2	2.50
Smoker (PY)	79	97.50
Family history of cancer	9	11.00
Occupational exposure		
None	72	88.90
Possible	9	11.10

Table 3. Gene ontology (GO) analysis of statistically significant up- and downregulated genes between control and NSCLC samples

Upregulated in cancer tissues			Downregulated in cancer tissues		
p value	GO term	biological process	p value	GO term	biological process
1.05E-29	GO:0022403	cell cycle phase	1.79E-13	GO:0009605	response to external stimulus
1.81E-29	GO:0000279	M phase	3.37E-13	GO:0032502	developmental process
1.11E-26	GO:0000278	mitotic cell cycle	2.11E-12	GO:0009611	response to wounding
1.43E-25	GO:0051301	cell division	9.36E-12	GO:0048856	anatomical structure development
1.04E-15	GO:0007051	spindle organization and biogenesis	2.25E-10	GO:0022610	biological adhesion
2.61E-13	GO:0006996	organelle organization and biogenesis	2.25E-10	GO:0007155	cell adhesion
8.18E-12	GO:0010564	regulation of cell cycle process	2.84E-10	GO:0048731	system development
2.03E-10	GO:0051276	chromosome organization and biogenesis	3.22E-10	GO:0007275	multicellular organismal development
7.01E-10	GO:0006259	DNA metabolic process	4.07E-10	GO:0032501	multicellular organismal process
3.77E-09	GO:0000226	microtubule cytoskeleton organization and biogenesis	6.13E-10	GO:0048513	organ development
3.82E-09	GO:0007059	chromosome segregation	2.41E-08	GO:0065008	regulation of biological quality
4.09E-09	GO:0007017	microtubule-based process	2.92E-08	GO:0035295	tube development
4.43E-09	GO:0006260	DNA replication	4.93E-08	GO:0009653	anatomical structure morphogenesis
1.76E-08	GO:0000070	mitotic sister chromatid segregation	5.99E-08	GO:0048869	cellular developmental process
2.50E-08	GO:0016043	cellular component organization and biogenesis	5.99E-08	GO:0030154	cell differentiation
3.83E-07	GO:0006323	DNA packaging	3.11E-07	GO:0006954	inflammatory response
5.29E-07	GO:0030261	chromosome condensation	7.47E-07	GO:0006952	defense response

estingly, the gene expression profile of the control samples was much more similar to lung adenocarcinomas than to squamous cell lung cancers.

GO Analysis of NSCLC

GO analysis of statistically significant up- and downregulated NSCLC-associated genes (1,103 and 672, respectively; $p \leq 0.05$ after Bonferroni's correction, fold change ≥ 2) suggested the involvement of wide-ranging cancer-related processes, including cell proliferation, tissue signaling, tissue connections, overall metabolism, and cell cycle regulation (table 3).

GO analysis of upregulated genes showed the activation of cell cycle regulation, mitosis, cell division, and DNA metabolism, suggesting a direct link between uncontrolled cell proliferation in NSCLC. Genes associated with cell adhesion, cell differentiation, inflammation, response to wounding, and defense responses were downregulated in cancer specimens. GO analysis of genes downregulated in NSCLC showed the involvement of much more diverse processes, including anatomical structure development, wound healing, and cell adhesion.

Fig. 1. Gene expression profiles of different histological subtypes of NSCLC. The lung adenocarcinoma samples are distinguishable from the squamous cell lung cancer samples. Adenocarcinomas and their subtype bronchioloalveolar carcinomas are clustered close to each other.

Cluster Analysis of Statistically Relevant Genes

We identified 997 statistically significant differentially expressed transcripts through our comparison of NSCLC versus control samples (326 upregulated genes and 671 downregulated genes). To visualize the expression profiles, we performed a hierarchical cluster analysis (fig. 2). According to the gene expression profiles, we identified 2 distinct cancer groups (referred to herein as group 1 and group 2). Based on their expression profiles, group 2 (blue) was more similar to the control samples and had a favorable survival prognosis. Although all of the control samples were clustered together, the different histologies were aligned randomly. According to our analysis, there was no clear association between the NSCLC gene expression profiles and NSCLC stage, smoking cessation, or patient gender.

Differentially Expressed Genes between NSCLC and Control Samples

The analysis of differentially expressed genes revealed a large number of previously described NSCLC-associated genes (online suppl. table 1; for supplementary material, see www.karger.com/doi/10.1159/000322116) as well as several potentially novel biomarkers of the disease. The group of novel upregulated genes associated with NSCLC included SPAG5, POLQ, KIF23, RAD54L, RAB26, and ARHGEF19, as well as 4 additional previously uncharacterized open reading frames. The group

Fig. 2. Gene expression heat map of NSCLC samples versus control samples. We used the TNM system to classify the NSCLC tumors according to the spread where stage 1 = Ia, stage 2 = Ib, stage 3 = IIa, stage 4 = IIb, stage 5 = IIIa, stage 6 = IIIb, and stage 7 = IV.

of novel downregulated genes associated with NSCLC included SGCG, NLRC4, VAPA, SFTPA1B, MMRN1, SFTPD, SELPLG, and PCDH17 (table 4). The analysis of the gene expression profiles of these novel genes revealed no association with the histological classification of the NSCLC.

Sperm-associated antigen 5 (also known as SPAG5, MAP126, and DEEPEST) encodes a protein associated with the functional and dynamic regulation of mitotic spindles [16]. It has been shown that the silencing of SPAG5 by RNAi in vitro results in growth arrest and the disruption of multipolar spindles, eventually leading to apoptosis [17]. Recent studies have also shown an interaction between SPAG5 with p53 [18] and kinetochores [19], showing that SPAG5 may play a key role in cell cycle regulation. Taken together with previous reports of SPAG5, our study found that SPAG5 is upregulated in NSCLC samples, suggesting that SPAG5 may be a new NSCLC biomarker with a role in cell cycle regulation and mitotic spindle organization.

Polymerase theta (also known as POLH, PRO0327, DKFZp781A0112, and POLQ) encodes a member of the Y family of specialized DNA polymerases. It copies undamaged DNA with a lower fidelity than other DNA-directed polymerases [20]. However, it accurately replicates UV-damaged DNA, and when thymine dimers are present, POLH incorporates complementary nucleotides into the newly synthesized DNA strand, thereby bypassing the lesion and suppressing the mutagenic effect of UV-induced DNA damage [21]. POLH is thought to be involved in hypermutation during immunoglobulin class switch recombination. Mutations in POLH result in xeroderma pigmentosum, variant type, which is a variant form of xeroderma pigmentosum [22, 23].

Although we are not aware of previous reports of POLH upregulation in cancer, one could hypothesize that there is a role for POLH in DNA metabolism, particularly under the extreme conditions present in fast-replicating cancer cells. Whether POLH upregulation in NSCLC may be induced by tobacco smoke carcinogens

Table 4. Potential novel NSCLC biomarkers

Adjusted p value	Average fold change	Gene symbol	Gene name and source
<i>Upregulated in cancer tissues</i>			
5.41E-16	2.2	C6ORF129	chromosome 6 open reading frame 129 [source:uniprot/sptrembl;acc:q5t7f7]
9.50E-14	3.6	SPAG5	sperm-associated antigen 5, map126, deepest [source:uniprot/swissprot;acc:q96r06]
1.11E-11	3.3	POLQ	dna polymerase theta [source:uniprot/swissprot;acc:o75417]
5.42E-11	2.1	C6ORF125	uncharacterized protein c6orf125 [source:uniprot/swissprot;acc:q9brt2]
3.13E-10	2.9	KIF23	kinesin-like protein kif23 [source:uniprot/swissprot;acc:q02241]
1.01E-09	2.2	RAD54L	DNA repair and recombination protein rad54-like [source:uniprot/swissprot;acc:q92698]
1.24E-09	2.2	C12ORF48	upf0419 protein c12orf48 [source:uniprot/swissprot;acc:q9nws1]
2.04E-09	2.0	C16ORF33	u11/u12 snrnp 25 kda protein (minus-99 protein) [source:uniprot/swissprot;acc:q9bv90]
4.38E-09	2.0	RAB26	ras-related protein rab-26 [source:uniprot/swissprot;acc:q9ulw5]
4.94E-09	2.0	ARHGEF19	rho guanine nucleotide exchange factor 19 [source:uniprot/swissprot;acc:q8iw93]
<i>Downregulated in cancer tissues</i>			
314E-20	2.4	SGCG	gamma-sarcoglycan [source:uniprot/swissprot;acc:q13326]
2.22E-17	3.5	NLRC4	caspase recruitment domain-containing protein 12 [source:uniprot/swissprot;acc:q9npp4]
1.75E-16	2.1	VAPA	vesicle-associated membrane protein-associated protein [source:uniprot/swissprot;acc:q9p0l0]
3.11E-15	9.5	SFTPA1B	pulmonary surfactant-associated protein a1 precursor [source:uniprot/swissprot;acc:q8iwl2]
2.65E-12	2.0	MMRN1	multimerin-1 precursor (endothelial cell multimerin 1) [source:uniprot/swissprot;acc:q13201]
3.52E-08	10.9	SFTPD	pulmonary surfactant-associated protein d precursor [source:uniprot/swissprot;acc:p35247]
6.44E-07	2.0	SELPLG	p-selectin glycoprotein ligand 1 precursor [source:uniprot/swissprot;acc:q14242]
7.99E-07	2.1	PCDH17	protocadherin-17 precursor (protocadherin-68) [source:uniprot/swissprot;acc:o14917]

and whether it may be responsible for the accumulation or elimination of mutations requires further analysis. Because previous studies suggest that POLH expression could be an important determinant of the cellular response to cisplatin [24], we propose that POLH should be added to the cisplatin marker set.

Kinesin family member 23 (also known as KIF23, CHO1, KNSL5, MKLP1, and MKLP-1) is a member of the kinesin-like protein family, which includes microtubule-dependent molecular motors that transport organelles within cells and move chromosomes during cell division [25]. KIF23 has been shown to cross-bridge antiparallel microtubules and drive microtubule movement in vitro. KIF23 interacts with Aurora family protein kinases, thereby regulating cell division and ensuring proper late-stage cytokinesis [26]. Because Aurora family proteins are known to be upregulated in various cancers and KIF23 was identified as an upregulated gene in the current study, we propose that KIF23 may play a role in NSCLC development.

The RAD54-like gene (also known as RAD54L, HR54, hHR54, RAD54A, and hRAD54) belongs to the superfamily of DEAD-like helicases and is proposed to be associated with homologous recombination and dsDNA break repair [27]. The binding of RAD54L to dsDNA induces a DNA topological change which stimulates DNA

recombination. Several mutations and polymorphisms in RAD54L have been associated with an elevated cancer risk and poor survival [28, 29]. However, little is known about RAD54L expression in cancer. We propose that RAD54L upregulation could be a natural cellular response to increased mutation rates in cancerous tissue.

We identified sarcoglycan gamma (also known as SGCG, DMDA, SCG3, TYPE, DAGA4, DMDA1, LGM-D2C, and SCARMD2) as a gene that is downregulated in NSCLC. SGCG encodes 1 of several sarcolemmal transmembrane glycoproteins that interact with dystrophin, thereby providing a structural link between the subsarcolemmal cytoskeleton and the extracellular matrix of muscle cells. Defects in the SGCG-encoded protein and downregulation of the SGCG gene can lead to the early onset of autosomal recessive muscular dystrophy, particularly limb-girdle muscular dystrophy type 2C [30]. Although we did not find any evidence in the literature of an association between SGCG and cancer, it is well known that both the extracellular matrix and the cytoskeleton are modified during cancer progression and metastasis. The downregulation and association of SGCG with NSCLC progression and metastatic potential require follow-up studies.

NLR family CARD domain-containing 4 (also known as NLRC4, CLAN, IPAF, and CARD12) is a gene acting between TP53 and caspase I that regulates cell cycle ar-

Fig. 3. Kaplan-Meier diagrams and prediction of NSCLC patient survival. Grouping of patients was based on the gene expression profiles of statistically significant up- and downregulated genes ($n = 997$). The survival curve obtained from the gene expression profiles of NSCLC patients yields a better survival prediction than histological grouping. AD/BA = Adenocarcinoma and bronchioloalveolar cancer; EPI = epidermoid cancer (squamous cell lung cancer.)

Fig. 4. Kaplan-Meier diagram of lung adenocarcinoma patient survival based on RNA integrity. There were 12 adenocarcinoma samples with a low RNA integrity ($RIN < 7$) and 21 samples with a high RNA integrity. Lung adenocarcinoma patients whose tumor specimens contained intact RNA had a statistically significantly higher survival prediction ($p = 0.0474$).

rest, apoptosis, and DNA repair [31]. NLRC4 is also proposed to induce apoptosis via NF- κ B signaling pathways [32]. NLRC4 has a caspase recruitment domain (CARD) through which it assembles into apoptosis and NF- κ B signaling complexes. The NLRC4 downregulation found in the current study should be monitored in subsequent studies as it is involved in several ways in the process of cancer initiation and progression.

Multimerin 1 (MMRN1), which we identified as a downregulated gene in NSCLC, is a massive, soluble protein found in platelets and in the endothelium of blood vessels. Multimerin is a factor V/Va-binding protein and may function as a carrier protein for platelet factor V [33]. Previous studies have demonstrated that MMRN1 interacts with factor V as well as with activated factor V with a high affinity, thereby inhibiting thrombin generation [34]. It has been suggested that MMRN1 could function as an adhesive ligand that promotes platelet adhesion at sites of vascular injury [35]. The downregulation of MMRN1 may contribute to leakage and the poor repair of blood vessels in cancer tissue, thereby allowing access of oxygen and nutrients to the cancer cell.

Surfactant protein D (also known as SFTPD and SP-D) is a member of the collagenous subfamily of calcium-dependent lectins (collectins). SFTPD is expressed in coronary artery smooth muscle cells as well as in pulmonary

alveolar type II cells where it modifies and modulates the inflammatory processes and host defense against pathogens [36, 37]. Recently, a study of surfactant protein D expression in a bronchioloalveolar lavage from a smoker revealed an association between reduced SFTPD expression and the progression of bronchial dysplasia [38]. From these limited but intriguing findings, we hypothesize that a decreased expression of SFTPD could be one of the early events leading to uncontrolled inflammation and cancer development.

These novel genes warrant further study in order to determine their NSCLC biomarker status, and to investigate their role in the initiation and progression of NSCLC.

Comparison of Survival Based on Gene Expression and Histology

As one of the key questions of the molecular diagnostics of cancer is survival prediction, we performed an analysis in which Kaplan-Meier diagrams were drawn based on gene expression profiles (group 1 and group 2) and histology. Although the time between the initial surgical resection and the survival analysis was limited to less than 7 years and the results of the analysis did not achieve statistical significance ($p = 0.0691$ for expression-based groups and $p = 0.0198$ for histology-based groups), we were able to show enhanced predictive p values in a group that was selected based on gene expression profiles (fig. 3).

Association between RNA Degradation and Survival

To investigate the hypothesis that RNA degradation in NSCLC specimens is associated with the prognosis of the disease, we performed a survival analysis. We did not identify a statistically significant association between NSCLC survival and RNA integrity status when comparing the lung cancer types as a whole. However, we did identify an association between RNA degradation and survival among patients with lung adenocarcinoma. The comparison of the survival of groups representing the intact RNA in the cancer sample and the degraded RNA showed a statistically significant distinction ($p = 0.0474$) indicating a much better survival in adenocarcinoma patients with intact RNA in cancer tissues (fig. 4).

Discussion

Although the prevalence and mortality of NSCLC are among the highest worldwide and the disease has been studied extensively, there is still a need for more accurate prognostic and diagnostic markers as well as for informa-

tion about the underlying molecular events in the disease's development. In the current gene expression study, we performed several analyses of NSCLC, including GO, clustering, new biomarker discovery, a comparison of survival based on histology and gene expression profiles, and analysis of RNA degradation status and patient survival.

In the cluster analyses of different histological subtypes of NSCLC, we could clearly distinguish adenocarcinoma and squamous cell cancer profiles. According to our results, the adenocarcinoma profile is more similar to the control than the squamous cell carcinoma profile. This phenomenon could be explained by the more prevalent cells of alveolar origin in the control tissue samples. In addition, there is evidence that lung adenocarcinoma patients have an increased prediction of survival compared to epidermoid lung cancer patients [39]. Although there are reports of different molecular profiles of adenocarcinomas [40, 41], we were not able to identify different adenocarcinoma profiles in our cohort. This might be due to the limited number of adenocarcinoma samples in our study group.

GO analysis of statistically significant upregulated genes in NSCLC uncovered biological processes involved in mitosis and cell cycle regulation, whereas GO analysis of statistically significant downregulated genes in NSCLC uncovered genes involved in much more diverse processes, such as anatomical structure development, wound healing, and cell adhesion.

The cluster analyses of statistically relevant deregulated genes between the NSCLC and control samples revealed 2 molecularly distinct patient groups with mixed histology, stage, and smoking history. The analyses of Kaplan-Meier diagrams based on gene expression profiles (group 1 and group 2) and histology showed enhanced predictive values ($p = 0.0198$) for gene expression-based comparison.

During the initial phase of the study, we eliminated a substantial number of cancer samples due to RNA degradation. Therefore, we also analyzed the relationship between survival in different histological groups and RNA status. Although we did not identify a statistically significant association between RNA integrity status and lung cancer in general, we did identify a statistically significant association in lung adenocarcinoma, showing that high levels of RNA degradation in patient samples were associated with decreased patient survival.

The novel NSCLC biomarkers brought to light in the current study and their applicability as therapy targets will require further validation with independent datasets in future studies.

Acknowledgements

This study was supported by Targeted Financing from the Estonian Ministry of Education and Research (SF0180142s08), Estonian Science Foundation grant ETF6465, EU FP7 grant ECOGENE (No. 205419, EBC), and by the EU via a European Regional Development Fund grant to the Centre of Excellence in Genomics, the Estonian Biocentre, and the University of Tartu.

Disclosure Statement

The authors declare no conflicts of interest.

References

- 1 Jemal A, et al: Cancer statistics, 2008. *CA Cancer J Clin* 2008;58:71–96.
- 2 Brambilla E, et al: The new World Health Organization classification of lung tumours. *Eur Respir J* 2001;18:1059–1068.
- 3 D'Amico TA: Molecular biologic staging of lung cancer. *Ann Thorac Surg* 2008;85:S737–S742.
- 4 Golub TR: Genome-wide views of cancer. *N Engl J Med* 2001;344:601–602.
- 5 Travis WD, Travis LB, Devesa SS: Lung cancer. *Cancer* 1995;75(suppl 1):191–202.
- 6 Mountain CF: The international system for staging lung cancer. *Semin Surg Oncol* 2000;18:106–115.
- 7 Smyth GK: Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 2004;3:article 3.
- 8 Reimand J, et al: g:Profiler – a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res* 2007;35(web server issue):W193–W200.
- 9 Petty RD, et al: Tumor transcriptome reveals the predictive and prognostic impact of lysosomal protease inhibitors in non-small-cell lung cancer. *J Clin Oncol* 2006;24:1729–1744.
- 10 Tsai YS, et al: RNA silencing of Cks1 induced G2/M arrest and apoptosis in human lung cancer cells. *IUBMB Life* 2005;57:583–589.
- 11 Bankovic J, et al: Identification of genes associated with non-small-cell lung cancer promotion and progression. *Lung Cancer* 2010;67:151–159.
- 12 Grinberg-Rashi H, et al: The expression of three genes in primary non-small cell lung cancer is associated with metastatic spread to the brain. *Clin Cancer Res* 2009;15:1755–1761.
- 13 Zhu YM, Bagstaff SM, Woll PJ: Production and upregulation of granulocyte chemotactic protein-2/CXCL6 by IL-1beta and hypoxia in small cell lung cancer. *Br J Cancer* 2006;94:1936–1941.
- 14 Ma Y, et al: Expression of targeting protein for xk1p2 associated with both malignant transformation of respiratory epithelium and progression of squamous cell lung cancer. *Clin Cancer Res* 2006;12:1121–1127.
- 15 Sienel W, et al: Melanoma associated antigen (MAGE)-A3 expression in stages I and II non-small cell lung cancer: results of a multicenter study. *Eur J Cardiothorac Surg* 2004;25:131–134.
- 16 Chang MS, et al: Cloning and characterization of hMAP126, a new member of mitotic spindle-associated proteins. *Biochem Biophys Res Commun* 2001;287:116–121.
- 17 Gruber J, et al: The mitotic-spindle-associated protein astrin is essential for progression through mitosis. *J Cell Sci* 2002;115(part 21):4053–4059.
- 18 Yang YC, et al: Silencing of astrin induces the p53-dependent apoptosis by suppression of HPV18 E6 expression and sensitizes cells to paclitaxel treatment in HeLa cells. *Biochem Biophys Res Commun* 2006;343:428–434.
- 19 Thein KH, et al: Astrin is required for the maintenance of sister chromatid cohesion and centrosome integrity. *J Cell Biol* 2007;178:345–354.
- 20 Matsuda T, et al: Error rate and specificity of human and murine DNA polymerase eta. *J Mol Biol* 2001;312:335–346.
- 21 Sary A, et al: Role of DNA polymerase eta in the UV mutation spectrum in human cells. *J Biol Chem* 2003;278:18767–18775.
- 22 Inui H, et al: Xeroderma pigmentosum-variant patients from America, Europe, and Asia. *J Invest Dermatol* 2008;128:2055–2068.
- 23 Masaki T, et al: Four types of possible founder mutations are responsible for 87% of Japanese patients with Xeroderma pigmentosum variant type. *J Dermatol Sci* 2008;52:144–148.
- 24 Albertella MR, et al: A role for polymerase eta in the cellular tolerance to cisplatin-induced damage. *Cancer Res* 2005;65:9799–9806.
- 25 Kuriyama R, et al: CHO1, a mammalian kinesin-like protein, interacts with F-actin and is involved in the terminal phase of cytokinesis. *J Cell Biol* 2002;156:783–790.
- 26 Neef R, et al: Cooperation between mitotic kinesins controls the late stages of cytokinesis. *Curr Biol* 2006;16:301–307.
- 27 Bugreev DV, Hanaoka F, Mazin AV: Rad54 dissociates homologous recombination intermediates by branch migration. *Nat Struct Mol Biol* 2007;14:746–753.
- 28 Bethke L, et al: Comprehensive analysis of DNA repair gene variants and risk of melanoma. *J Natl Cancer Inst* 2008;100:270–276.
- 29 Li D, et al: DNA repair gene polymorphisms and risk of pancreatic cancer. *Clin Cancer Res* 2009;15:740–746.
- 30 Klinge L, et al: Sarcoglycanopathies: can muscle immunoanalysis predict the genotype? *Neuromuscul Disord* 2008;18:934–941.
- 31 Sadasivam S, et al: Caspase-1 activator Ipaf is a p53-inducible gene involved in apoptosis. *Oncogene* 2005;24:627–636.
- 32 Geddes BJ, et al: Human CARD12 is a novel CED4/Apaf-1 family member that induces apoptosis. *Biochem Biophys Res Commun* 2001;284:77–82.
- 33 Jeimy SB, et al: Identification of the MMRN1 binding region within the C2 domain of human factor V. *J Biol Chem* 2004;279:51466–51471.
- 34 Jeimy SB, et al: Multimerin 1 binds factor V and activated factor V with high affinity and inhibits thrombin generation. *Thromb Haemost* 2008;100:1058–1067.
- 35 Tasneem S, et al: Platelet adhesion to multimerin 1 in vitro: influences of platelet membrane receptors, von Willebrand factor and shear. *J Thromb Haemost* 2009;7:685–692.
- 36 Hilgendorff A, et al: Association of polymorphisms in the human surfactant protein-D (SFTPD) gene and postnatal pulmonary adaptation in the preterm infant. *Acta Paediatr* 2009;98:112–117.
- 37 Snyder GD, et al: Surfactant protein D is expressed and modulates inflammatory responses in human coronary artery smooth muscle cells. *Am J Physiol Heart Circ Physiol* 2008;294:H2053–H2059.
- 38 Sin DD, et al: Surfactant protein D and bronchial dysplasia in smokers at high risk of lung cancer. *Chest* 2008;134:582–588.
- 39 Merrill RM, Henson DE, Barnes M: Conditional survival among patients with carcinoma of the lung. *Chest* 1999;116:697–703.
- 40 Bhattacharjee A, et al: Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci USA* 2001;98:13790–13795.
- 41 Beer DG, et al: Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med* 2002;8:816–824.