Carcinogenesis vol.34 no.12 pp.2767–2773, 2013
doi:10.1093/carcin/bgt294
Advance Access publication August 25, 2013

Gene expression signature of non-involved lung tissue associated with survival in lung adenocarcinoma patients

Antonella Galvan[1,†], Elisa Frullanti[1,8,†], Marco Anderlini[1], Giacomo Manenti[1], Sara Noci[1], Matteo Dugo[1], Federico Ambrogi[2], Loris De Cecco[1], Roberta Spinelli[3], Rocco Piazza[3], Alessandra Pirola[3], Carlo Gambacorti-Passerini[3,4], Matteo Incarbone[5,6], Marco Alloisio[5], Davide Tosi[7], Mario Nosotti[7], Luigi Santambrogio[7], Ugo Pastorino[1] and Tommaso A.Dragani[1,*]

[1]Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy, [2]Department of Clinical Sciences and Community Health, University of Milan, Milan, Italy, [3]Department of Health Sciences, University of Milano-Bicocca, Monza, Italy, [4]Hematology and Clinical Research Unit, San Gerardo Hospital, Monza, Italy, [5]Department of Surgery, Istituto Clinico Humanitas, Rozzano, Italy, [6]Department of Surgery, San Giuseppe Hospital, Multimedica, Milan, Italy and [7]Department of Surgery, IRCCS Fondazione Cà Granda Ospedale Maggiore Policlinico, Università degli Studi di Milano, Milan, Italy
[8]Present address: Medical Genetics, University of Siena, Siena, Italy

*To whom correspondence should be addressed. Department of Predictive and Preventive Medicine, Fondazione IRCCS Istituto Nazionale dei Tumori, Via Amadeo 42, I-20133 Milan, Italy. Tel: +39 0223902642; Fax: +39 0223902764;
Email: tommaso.dragani@istitutotumori.mi.it

Lung adenocarcinoma patients of similar clinical stage and undergoing the same treatments often have marked interindividual variations in prognosis. These clinical discrepancies may be due to the genetic background modulating an individual's predisposition to fighting cancer. Herein, we hypothesized that the lung microenvironment, as reflected by its expression profile, may affect lung adenocarcinoma patients' survival. The transcriptome of non-involved lung tissue, excised from a discovery series of 204 lung adenocarcinoma patients, was evaluated using whole-genome expression microarrays (with probes corresponding to 28 688 well-annotated coding sequences). Genes associated with survival status at 60 months were identified by Cox regression analysis (adjusted for gender, age and clinical stage) and retested in a validation series of 78 additional cases. RNA-Seq analysis from non-involved lung tissue of 12 patients was performed to characterize the different isoforms of candidate genes. Ten genes for which the $\log_e$-transformed hazard ratios expressed the same direction of effect in the discovery ($P < 1.0 \times 10^{-3}$) and validation series comprised the gene expression signature associated with survival: *CNTNAP1*, *PKNOX1*, *FAM156A*, *FRMD8*, *GALNTL1*, *TXNDC12*, *SNTB1*, *PPP3R1*, *SNX10* and *SERPINH1*. RNA sequencing highlighted the complex expression pattern of these genes in non-involved lung tissue from different patients and permitted the detection of a read-through gene fusion between *PPP3R1* and the flanking gene (*CNRIP1*) as well as a novel isoform of *CNTNAP1*. Our findings support the hypothesis that individual genetic characteristics, evidenced by the expression pattern of non-involved tissue, influence the outcome of lung adenocarcinoma patients.

Introduction

Establishing the prognosis of a patient with a particular primary cancer is difficult because of the marked interindividual variations in tumor progression and response to treatment. Thus, patients with apparently similar cancer characteristics may have substantially different clinical outcomes for unknown reasons, justifying active research on prognostic factors (1,2). This situation is particularly true for lung cancer, a heterogeneous disease for which a greater genetic understanding would improve our ability to predict survival and deliver effective treatments.

Numerous recent studies have searched for genetic variations that associate with survival in different types of lung cancer. For example, in advanced lung cancers, some studies have investigated the relationship between somatic mutations in specific oncogenes and survival (3,4). Preliminary studies in non-small-cell lung cancer have detected germ line variations represented by single nucleotide polymorphisms that modulate survival (5,6). Interestingly, germ line variations may modulate the expression of other genes, thus defining expression quantitative trait loci (7). In lung adenocarcinoma—a common histological type of lung cancer with increasing incidence (8)—two recent studies identified tumoral gene expression profiles having prognostic information (9,10). Also working in lung adenocarcinoma, we identified germ line single nucleotide polymorphisms associating with clinical stage (11) and reported an association between the transcriptional profile of non-involved lung tissue and the clinical stage of the nearby tumor (12); these preliminary findings support the hypothesis that genetic constitution may modulate clinically relevant parameters in lung adenocarcinoma patients. In addition, results from experimental models showing differences in lung adenoma/adenocarcinoma size in different strains of mice suggest that germ line differences may control lung tumor aggressiveness (13).

Herein, we analyzed the transcriptome of non-involved lung tissue in lung adenocarcinoma patients to assess whether or not individual differences in gene expression may represent predictors of survival.

Materials and methods

Study population and tissue samples

The study employed a biobank of samples of non-involved (apparently normal) lung parenchyma excised from patients who underwent lobectomy for lung adenocarcinoma in the authors' institutes in the area around Milan, Italy. At the end of surgery, a small section of non-involved tissue was taken from the excised lobe as far as possible from the cancer tissue; it was stored frozen or placed directly in RNAlater solution (Life Technologies, Grand Island, NE). Part of these samples has been used in a study on clinical stage (12). In some cases, the non-involved lung tissue was matched with a specimen of lung adenocarcinoma tissue. Information on histological diagnoses (made by the Pathology Departments of the recruiting institute or hospital) was retrieved from the clinical records. Data regarding gender, age at diagnosis and clinical stage were recorded when the samples were taken. Survival status after surgery was also recorded; we did not consider survival after 60 months to avoid possible bias due to deaths not related to cancer.

For the purposes of this study, we used 282 samples of non-involved tissue from ever-smokers; this methodological choice allowed us to avoid possible bias in gene expression associated with smoking habit (14). The samples were analyzed in two sets: a discovery series ($n = 204$) and a validation series that became available after the discovery series was analyzed ($n = 78$). The study protocol was approved by the Committees for Ethics of the institutes involved in recruitment (Fondazione IRCCS Istituto Nazionale dei Tumori, Istituto Clinico Humanitas, San Giuseppe Hospital, Ospedale Maggiore Policlinico). Each patient gave informed written consent to the use of their biological samples and data for research purposes.

RNA extraction and gene expression analysis

At the Istituto Nazionale dei Tumori in Milan, total RNA was extracted from the lung tissue samples using Trizol reagent (Life Technologies) following the manufacturer's instructions, treated with DNase I (Qiagen, Santa Clarita, CA) and quantified by spectrophotometry (ND-2000c; NanoDrop Products, Wilmington, DE). RNA integrity was verified using the RNA 6000 Nano Kit (Agilent Technologies, Palo Alto, CA); all samples had an RNA integrity number >7, indicating good quality. RNA was reverse transcribed, labeled with biotin and amplified overnight using the Illumina TotalPrep RNA Amplification Kit (Life Technologies).

Abbreviations: cDNA, complementary DNA; HR, hazard ratio; mRNA, messenger RNA.

[†]These authors contributed equally to the work.

Biotinylated complementary RNA (1.5 µg per sample) was diluted in E1 hybridization buffer and hybridized to HumanHT-12 v4 Expression BeadChips (Illumina, San Diego, CA). These microarrays contain over 47 000 probes, including 28 688 probes that correspond to well-annotated coding sequences. Hybridization was done first for all samples of the discovery series and, subsequently, for the validation series. After hybridization, the BeadChips were washed following the manufacturer's protocol and scanned with an Illumina BeadArray Reader. Primary data were collected using BeadStudio v3 software package. After quality control, microarray data were $\log_2$ transformed and normalized using the robust spline normalization method, implemented in the lumi package (15) of the open source software Bioconductor (16).

When multiple probes represented the same transcript, we included only the one with the highest detection rate, defined as the percentage of samples in which the probe had a detection $P < 0.01$ (this P value represents the confidence that a given transcript is expressed above the background level defined by negative control probes). Probes that were not annotated were also eliminated from analysis. Finally, in the case of the discovery set, we filtered the remaining probes and kept only those with a detection $P < 0.01$ in at least 90% of samples. In the validation set, we kept all probes for which a detection $P < 0.01$ was obtained for at least one sample; these less stringent criteria reduced the possibility that genes identified in the discovery series be absent from the validation set. Log_2-transformed and normalized values of the resulting transcripts were then used in Cox proportional hazards modeling.

High-throughput RNA sequencing

High-throughput RNA sequencing (RNA-Seq) was carried out in order to identify isoforms differentially expressed in non-involved lung tissue. This analysis was performed using 12 randomly selected samples from the discovery series (including one sample for which we also had a matched sample of lung adenocarcinoma). This subset included eight males and four females, seven with stage I and five with stage >I and eight who were alive at the 60 months of follow-up.

Messenger RNA (mRNA) was isolated from 2 µg total RNA using oligo-dT magnetic beads and it was fragmented at 94°C for 1 min and then prepared for sequencing according to the protocol of the TruSeq RNA Sample Prep Kit v2 (Illumina) with one additional step, namely the selection of 400–500 bp fragments on 2% agarose gels after the ligation of the adapters. The resulting complementary DNA (cDNA) libraries were sequenced on an Illumina Genome Analyzer IIx with 76 bp paired-end reads using Illumina TruSeq SBS kit v5.

Image processing and base calling were performed using the Illumina Real Time Analysis Software RTA v1.9.35. Qseq files were deindexed and converted to the Sanger-FastQ file format, using in-house scripts. FastQ sequences were aligned to the human genome database (NCBI36/hg18) using TopHat v.1.2.0 (17) with default parameters. The reads were mapped using the gene and splice junction models as provided in the annotation GTF ('Gene Transfer Format) file (Ensembl release 54). A splice junction map of each gene was inferred from TopHat and checked using the Integrated Genomic Viewer (18) and the Ensembl database (release 71, genome assembly: GRCh37).

Detection of gene fusions in pairs of non-involved lung tissue and lung adenocarcinoma

From our biobank, we obtained 41 pairs of matched non-involved lung tissue and lung adenocarcinoma: in 10 cases (including one of the 12 samples analyzed by RNA-Seq), the non-involved lung tissue had been used in the gene expression experiments. RNA from these pairs (1 µg per sample) was used to synthesize cDNA by reverse transcription using the Transcriptor First Strand cDNA Synthesis Kit (Roche, Basel, Switzerland) according to the manufacturer's instructions. To confirm the presence of the gene fusion between *PPP3R1* (NM_000945) and *CNRIP1* (NM_015463) using a different methodology, we first performed PCR on the sample analyzed by RNA-Seq using primer pairs designed to amplify a fragment of 357 bp (forward primer, 5′-ctgaagtctaaagagcctgatgg-3′; reverse primer, 5′-gaactgagagacgccctcaatg-3′) and another of 169 bp (forward primer, 5′-cccaacgaagagtggagaacg-3′; reverse primer, 5′-ctcttccacactcaaagaaccag-3′) containing the fusion. Then, on the remaining 40 samples, we PCR amplified the 357 bp fragment.

PCR reactions were carried out using 5 µl cDNA template (diluted 1:10), 0.2 µM primers and 0.5 U AmpliTaq Gold DNA Polymerase (Life Technologies) in a final volume of 25 µl. PCR-amplified fragments were visualized in 3% agarose gels where the DNA molecular weight marker was φX174 DNA-Hae III Digest (New England Biolabs).

Statistical analyses

Survival curves were estimated using the Kaplan–Meier method for all patients in the discovery or validation series and, in the discovery series, for patients distinguished into two groups according to whether their mRNA levels for particular genes were above or below the group median. The association of expression levels of individual genes with survival status at 60 months was evaluated through a Cox proportional hazard model adjusted for gender, age and clinical stage (stage I versus >I). The analysis was performed in three consecutive steps. First, using discovery series data, hazard ratios (HRs) were calculated for all genes and $\log_e$-transformed; genes were ranked according to the P values of the $\log_e$HR, and those with $P < 1.0 \times 10^{-3}$ were selected. Then, for these genes, $\log_e$HR and P values were computed using validation series data. Given that the validation series was smaller, we did not expect it to have statistical power to confirm the findings from the discovery series: thus, to identify a gene expression signature associated with survival, we considered those genes for which the direction of the effect (i.e. the sign of $\log_e$HR) was the same in the discovery and validation series. This criterion has been reported to represent a suitable validation method for use in population-based studies (19).

For those genes deemed potentially associated with survival, an overall $\log_e$HR and P value were computed by combining the discovery and validation series results according to standard meta-analysis procedures (20), performed using the metagen function of the R package meta. In particular, the $\log_e$HR were combined using a fixed effect or random effect model according to the result of the test of homogeneity. In addition, for these genes, the expression data were analyzed by hierarchical clustering using Pearson's correlation distance and average linkage. Gene expression data were analyzed with R software (http://www.R-project.org/). All P values were two sided.

Results

The study employed 282 samples of non-involved lung tissue from patients with lung adenocarcinoma, divided into a discovery series and a validation series (Table I). The two series were similar for age at cancer diagnosis and clinical stage distribution, although the discovery series had a higher proportion of women. Analysis of the survival curves of the two series did not show a significant difference (log-rank test $P = 0.138$; Supplementary Figure 1, available at *Carcinogenesis* Online). Therefore, the validation series was deemed suitable for confirming the results from the discovery series, despite its smaller size.

In both series, age at diagnosis and gender were not significantly associated with survival (data not shown). Instead, there was a strong, inverse association between clinical stage (stage I versus >I) and survival status at 60 months, as would be expected.

Association of lung tissue gene expression with overall survival

HumanHT-12 v4 Expression BeadChips were used to profile the transcriptome of non-involved lung tissue in the discovery series. After quality control filtering, data were available regarding the expression of 11 420 unique transcripts. Using a Cox proportional hazard model adjusted for gender, age and clinical stage, we observed that mRNA expression levels of 17 genes associated with overall survival at nominal $P < 1.0 \times 10^{-3}$ (Table II). HRs of these genes ranged from 0.2 to 6.5 ($\log_e$HR from −1.81 to 1.88). Seven genes had a $\log_e$HR <0, indicating that as mRNA levels increase the risk of death decreases

Table I. Phenotypic characteristics of lung adenocarcinoma patients, by study group

Characteristic	Discovery series ($n = 204$)	Validation series ($n = 78$)
Median age (range), years[a]	66 (36–83)	67 (49–85)
Gender, *n* (%)		
Male	138 (67.6)	62 (79.5)
Female	66 (32.4)	16 (20.5)
Clinical stage, *n* (%)[b]		
I	101 (50.2)	39 (50.0)
II	32 (15.9)	13 (16.6)
III	57 (28.4)	23 (29.5)
IV	11 (5.5)	2 (2.6)
Survival status at 60 months		
Alive	118	54
Dead	86	24

All patients were ever-smokers.
[a]Age at diagnosis.
[b]Data missing for three patients in the discovery series and one patient in the validation series.

(increased overall survival), whereas the other 10 genes had $\log_e$HR >0. The gene with the smallest *P* value was LIN54 [lin-54 homolog (*Caenorhabditis elegans*)] ($P = 3.4 \times 10^{-5}$; HR = 0.2; $\log_e$HR = −1.81).

The validation series, filtered using less stringent criteria, provided expression data regarding 16 103 transcripts. After adjusting for gender, age and clinical stage, none of the 17 genes identified in the discovery series reached nominal significance ($P < 0.05$) for the association with survival (Table II). Since the validation series was substantially smaller than the discovery series, these results were expected. Therefore, as an alternative indicator of an association with survival, we used the criterion of a similar direction of effect ($\log_e$HR ≥ zero) in both the discovery and validation series. This approach identified 10 genes potentially associated with survival: *CNTNAP1*, *PKNOX1*, *FAM156A*, *FRMD8*, *GALNTL1*, *TXNDC12*, *SNTB1*, *PPP3R1*, *SNX10* and *SERPINH1*. In the meta-analysis, these genes remained significantly associated with survival with *P* ranging from 2.2×10^{-5} to 3.2×10^{-3}. Five genes had positive $\log_e$HR values, whereas the other five had negative $\log_e$HR values; a positive $\log_e$HR value indicates a direct association between mRNA levels and risk of death, whereas a negative $\log_e$HR value indicates an inverse association.

Hierarchical clustering of transcript levels for these 10 genes in the discovery series revealed that they cluster into two distinct groups (Figure 1). In any given patient, when one gene cluster (*CNTNAP1*, *PKNOX1*, *FRMD8*, *GALNTL1* and *SERPINH1*) showed high expression levels, the other one (*FAM156A*, *TXNDC12*, *SNTB1*, *PPP3R1* and *SNX10*) usually showed low expression levels. The expression patterns of these 10 genes grouped the patients in different clusters, some of which were enriched with patients who had died before 60 months of follow-up.

To visualize differences in survival rate associated with differences in the expression of a particular gene, we divided patients of the discovery series into two equal groups on the basis of their expression level being either above or below the group median. Kaplan–Meier survival curves showed, in all cases, clear differences in survival between the high- and low-expression groups (Figure 2). These

Table II. Genes expressed in non-involved lung tissue of patients with lung adenocarcinoma and found to be associated with overall survival at 60 months by Cox proportional hazards regression analysis (with $P < 1.0 \times 10^{-3}$) in the discovery series, and Cox regression results for the validation series and meta-analysis

Symbol[a]	Discovery series ($n = 204$)		Validation series ($n = 78$)		Meta-analysis[c]		
	$\log_e$HR[b]	*P*	$\log_e$HR	*P*	$\log_e$HR (95% CI)	*P*	I^2
LIN54	−1.81	3.4E-05	1.39	2.8E-01			
CNTNAP1	1.00	4.1E-05	0.73	2.5E-01	0.96 (0.52–1.40)	2.2E-05	0.00
TP53BP1	1.65	7.4E-05	−0.51	6.0E-01			
THBS3	1.13	8.8E-05	−0.13	8.5E-01			
PKNOX1	1.87	1.5E-04	0.81	5.6E-01	1.75 (0.84–2.67)	1.7E-04	0.00
FAM156A	−1.51	1.8E-04	−0.94	2.2E-01	−1.38 (−2.08 to −0.69)	1.0E-04	0.00
FRMD8	1.19	1.9E-04	0.49	5.0E-01	1.08 (0.51–1.66)	2.3E-04	0.00
GALNTL1	0.80	2.5E-04	0.47	4.2E-01	0.76 (0.36–1.17)	2.1E-04	0.00
PPT2	1.88	3.3E-04	−1.53	3.1E-01			
TXNDC12	−1.41	3.7E-04	−0.26	7.8E-01	−1.23 (−1.94 to −0.52)	7.2E-04	0.24
FAM131A	1.25	5.4E-04	−0.51	6.4E-01			
SNTB1	−1.03	6.2E-04	−0.75	3.0E-01	−0.99 (−1.54 to −0.45)	3.7E-04	0.00
PPP3R1	−0.75	7.1E-04	−0.55	5.0E-01	−0.74 (−1.16 to −0.32)	5.8E-04	0.00
SNX10	−0.54	7.3E-04	−0.18	6.1E-01	−0.48 (−0.77 to −0.19)	1.0E-03	0.00
CDCP1	−0.88	7.4E-04	0.45	5.5E-01			
FUS	1.69	7.7E-04	−0.55	7.2E-01			
SERPINH1	0.79	8.8E-04	0.01	9.8E-01	0.62 (0.21–1.03)	3.2E-03	0.57

CI, confidence interval.
[a]Genes are listed in order of increasing *P* value in the discovery series. In bold type are genes whose direction of effects was maintained in the validation series and for which, the meta-analysis was carried out.
[b]Natural logarithm of the HR, obtained by Cox regression adjusted by gender, age at diagnosis and clinical stage (I versus >I).
[c]Meta-analysis was carried out for the 10 genes that passed validation. I^2 indicates heterogeneity between the discovery and validation series, i.e. it approximates the proportion of the total variation due to heterogeneity.

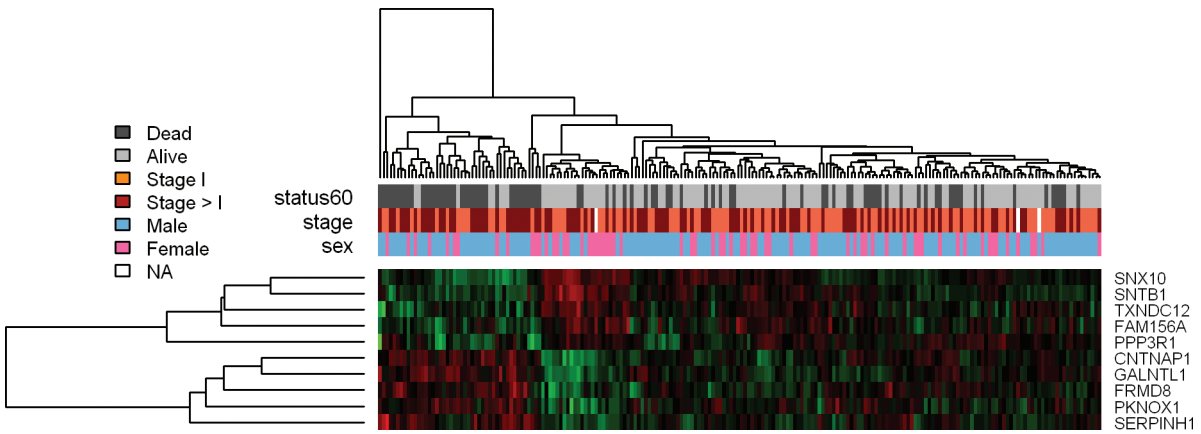

Fig. 1. Hierarchical clustering of the expression levels of the 10 genes expressed in non-involved lung tissue and associated with overall survival in patients with lung adenocarcinoma. Data refer to the 204 patients in the discovery series and are presented in a false-color scale where red means higher expression and green lower expression.

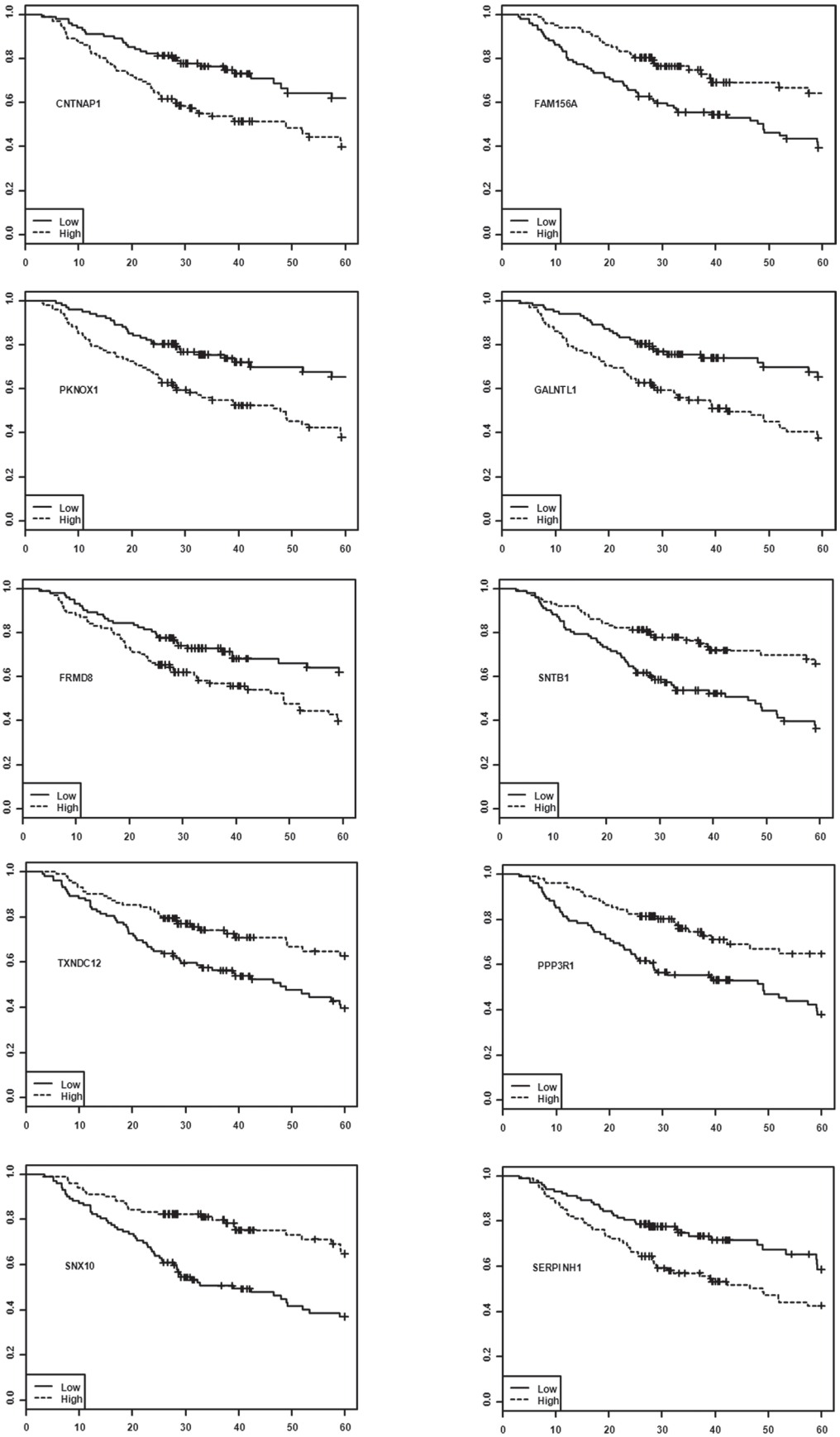

Fig. 2. Kaplan–Meier survival curves for 204 lung adenocarcinoma patients (discovery series), subdivided according to whether their mRNA levels were above or below the group median, for 10 genes associated with survival. Dashed gray and solid black lines indicate high- and low-expression level groups. Crosses represent censored samples. Values reported on the *x*-axes are months of follow-up, whereas values on the *y*-axes are probability of survival.

analyses confirm the results obtained with Cox modeling (Table II). Indeed, the groups with high expression levels (Figure 2, dashed gray lines) had worse survival for genes with a positive $\log_e$HR (*CNTNAP1*, *PKNOX1*, *FRMD8*, *GALNTL1*, *SERPINH1*), whereas they had better survival for genes with a negative $\log_e$HR (FAM156A, TXNDC12, SNTB1, PPP3R1, SNX10).

Main mRNA isoforms of the 10 genes associated with survival

To analyze the expression patterns of the ten genes potentially associated with cancer survival and to perform isoform analyses, we generated RNA-Seq data using 12 samples of non-involved lung tissue from the discovery series. According to Ensembl database (release 71), these genes have from 4 to 20 known (annotated) transcripts (Table III). Sequencing confirmed that all 10 genes are expressed in lung tissue. They are present as a main isoform in all samples and from zero (*SNTB1*) to six (*FAM156A*) minor isoforms in a subset of cases. In particular, we identified a novel isoform of *CNTNAP1* in six samples. This isoform is generated by the alternative splicing of the main isoform which eliminates exon 10 and produces a novel out-of-frame transcript not yet reported in public databases. Moreover, we found junction reads between the 5′ region of *PPP3R1* and the 3′-end of the flanking gene, *CNRIP1*, in three samples. These junction reads resulted from two different types of gene fusion: one involved a read-through between exon 2 of the *PPP3R1* gene and exon 2 of the *CNRIP1* gene (seen in one sample), whereas the other involved intronic regions of the genes (in two samples).

For the first fusion type, eight junction reads were detected in one of the samples. To confirm that this gene fusion existed in the patient and was not an experimental artifact, we used PCR to amplify two fragments around the fusion site. This analysis showed that the *PPP3R1–CNRIP1* gene fusion was indeed present in the non-involved lung tissue, and it was also present in the lung adenocarcinoma tissue, which was available from this patient (Figure 3). PCR amplification of the large PCR fragment from an additional 40 pairs of non-involved lung tissue and lung adenocarcinoma revealed that in 8 cases, the gene fusion fragment was detected in both samples from the same patient, in 3 cases, it was found only in the non-involved tissue, in 17 cases, it was only in the tumoral tissue, whereas in the remaining 13 cases, it was not detectable. Based on the intensity of the amplified bands, there was a tendency to higher expression levels in tumor samples than in non-involved tissue (Supplementary Figure 2, available at *Carcinogenesis* Online).

Discussion

We analyzed the transcriptome profile of non-involved lung tissue excised from patients with lung adenocarcinoma, with the aim of identifying a gene expression signature associated with patients'

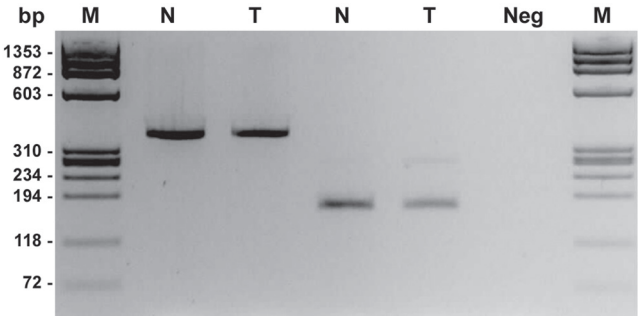

Fig. 3. Gene fusion between *PPP3R1* and *CNRIP1* genes. PCR amplification of PPP3R1–CNRIP1 fusion fragments in the patient in whom the gene fusion was first detected by RNA-Seq. PCR amplification in non-involved (N) or lung adenocarcinoma (T) tissue, using PCR primers amplifying either a long fragment (first two bands from the left) or a short fragment (third and fourth bands from the left). M, DNA molecular weight marker, φX174 DNA-Hae III Digest; Neg, negative control using water instead of DNA as PCR template.

Table III. Transcript isoforms of the 10 genes expressed in non-involved lung tissue and associated with cancer survival

Symbol[a]	No. of known transcripts[a]	Expressed transcripts	No. of samples in which the transcript was detected (%)[b]	Comments
CNTNAP1	4	ENST00000264638	12 (100)	
		ENST00000586801	10 (83.3)	Intron retention between exons 7 and 8
		Novel	6 (50.0)	Exon 10 skipped
PKNOX1	10	ENST00000291547	12 (100)	
		ENST00000432907	5 (41.7)	Exon 4 skipped
FAM156A	20	ENST00000356333	12 (100)	
		ENST00000438003	10 (83.3)	
		ENST00000316310	9 (75)	
		ENST00000330025	6 (50.0)	
		ENST00000505988	4 (33.3)	Not distinguishable from ENST00000414076
		ENST00000414076	4 (33.3)	Not distinguishable from ENST00000505988
		ENST00000512364	1 (8.3)	
FRMD8	9	ENST00000317568	12 (100)	
		ENST00000355991	4 (33.3)	Exon 3 skipped
		ENST00000416776	2 (16.7)	Exon 4 skipped
GALNTL1	10	ENST00000448469	12 (100)	
		ENST00000337827	12 (100)	
TXNDC12	4	ENST00000371626	12 (100)	
		ENST00000471493	2 (16.7)	Intron retention between exons 6 and 7
SNTB1	5	ENST00000517992	12 (100)	
PPP3R1	4	ENST00000234310	12 (100)	
		Novel	3 (25)	Gene fusion with CNRIP1
SNX10	8	ENST00000338523	12 (100)	
		ENST00000396376	8 (66.7)	
		ENST00000416246	7 (58.3)	Not distinguishable from ENST00000446848
		ENST00000446848	7 (58.3)	Not distinguishable from ENST00000416246
SERPINH1	15	ENST00000358171	12 (100)	
		ENST00000533603	12 (100)	

[a]Ensembl database (release 70).
[b]Total number of samples assayed = 12.

overall survival. Our analysis in a discovery series and a validation series led to the identification of 10 genes, half of which have an inverse association between mRNA expression levels and the risk of death. Besides being expressed in non-involved lung tissue, these 10 genes are also expressed in a range of tissues and cells, including inflammatory cells, immune cells and fibroblasts, as indicated by the GeneCards (http://www.genecards.org/) and the GEO (Gene Expression Omnibus) Profiles (http://www.ncbi.nlm.nih.gov/geoprofiles) databases. Because these cell types may also be found in the lung microenvironment, these 10 candidate genes may have a role in influencing the progression of the surrounding tumor cells.

mRNA sequencing confirmed the expression of all 10 genes in non-involved lung tissue and revealed a complex pattern of expression of their isoforms, with individual differences in the type of detectable isoforms. These results support the hypothesis that the genetic constitution modulates clinically relevant parameters in lung adenocarcinoma patients. Moreover, they indicate that the analysis of gene expression levels, without taking into consideration their specific transcript isoforms and relative expressions, may provide only partial information on the complex transcriptome profile of specific tissues. Therefore, RNA-Seq analysis of non-involved lung tissue from a large series of lung adenocarcinoma patients would provide novel data on gene transcript isoforms that may be associated with patients' outcome. Here, by RNA-Seq analysis, we detected a fusion between the *PPP3R1* and *CNRIP1* genes in non-involved lung tissue and observed that it is common in both non-involved tissue and lung adenocarcinoma tissue. Although we have no evidence of tumor contamination of the non-involved tissue, we cannot exclude that the fusion originated in the tumor cells or that it is due to a field cancerization effect (21), given its higher frequency and expression levels in the tumors than in the non-involved tissues. Additional studies are necessary to characterize this gene fusion (and other possible fusions) in non-involved lung in order to understand the role of such events in modulating lung cancer patients' survival.

To begin to understand how the 10 gene expression signature identified in this study may influence survival in patients with lung adenocarcinoma, information on the functions of the encoded proteins was obtained from the GeneCards database (http://www.genecards.org/) and from the cited literature therein (Table IV). Among these genes, *PKNOX1* has been recognized as a tumor suppressor gene involved in maintaining genomic stability (25,26), whereas the products of *FRMD8*, *GALNTL1* and *TXNDC12* belong to families of proteins whose members have been implicated in cancer aggressiveness or progression (22,24,34,35). Considering the results of this study and the fact that several of these proteins (or their related family members) have already been implicated in cancer biology, further research on their potential functional role in modulating the survival of lung adenocarcinoma patients is warranted.

In evaluating the findings from this study, it should be taken into account that the validation series was smaller than the discovery series, limiting our ability to statistically validate the association of the genes with survival. During the study period, we were unable to collect a larger number of samples despite the collaboration of surgeons at several hospitals in our region. Therefore, the validation analysis had to consider whether or not the trend of effects observed in the discovery series was reproduced in the second, independent series, as suggested by Zeggini *et al*. (19). Difficulties in validating results in independent series have also been reported in studies on the association of somatic transcriptional profiles of lung adenocarcinoma with survival rate (10,36,37). Recently, in lung adenocarcinoma cohorts, a 193 gene expression signature for tumor tissue was reported to associate with overall survival (10). However, this signature does not overlap, even partially, with other prognostic signatures, e.g. an 82 transcript expression signature in lung adenocarcinoma patients (38) or with another 51 gene expression signature in patients with non-small-cell lung carcinoma (39). Therefore, these profiles have not yet found clinical application in predicting survival for lung adenocarcinoma patients.

Differently from these previously mentioned studies, our study aimed to assess if the genetic constitution, leading to differences in gene expression in non-involved lung tissue, modulates patients' survival independently of somatic changes in their tumors. Our findings

Table IV. Proteins encoded by 10 genes associated with lung cancer survival

Symbol	Protein names	Description	References
CNTNAP1	Contactin-associated protein 1	Transcribed predominantly in brain and weakly expressed in other tissues, including lung.	http://www.genecards.org/
FAM156A	Family with sequence similarity 156, member A	Transmembrane protein of unknown function, ubiquitously expressed.	http://www.genecards.org/
FRMD8	FERM domain containing 8	Belongs to the FERM family of proteins, which are involved in tumor progression: FRMD5 knockdown promotes lung cancer cell migration and invasion *in vitro*, whereas FRMD4A upregulation correlates with increased risk of relapse in primary human head and neck squamous cell carcinoma.	(22,23)
GALNTL1	UDP-*N*-acetyl-α-D-galactosamine:polypeptide *N*-acetylgalactosaminyltransferase-like 1	A member of its family, GALNTL2, modifies the activity of the epidermal growth factor receptor, thus modulating tumor aggressiveness in hepatocellular carcinoma.	(24)
PKNOX1	PBX/knotted 1 homeobox 1; PREP1	Homeodomain transcription factor, ubiquitously expressed; involved in maintaining genomic stability; a candidate tumor suppressor gene.	(25,26)
PPP3R1	Protein phosphatase 3, regulatory subunit B, alpha	Regulatory subunit of calcineurin, modulating calcium sensitivity. Inactivation of Ppp3r1 causes lethal cardiomyopathy and lethal diabetes in mice.	(27,28)
SERPINH1	Serpin peptidase inhibitor, clade H (heat shock protein 47), member 1	Belongs to the serpin superfamily of serine proteinase inhibitors and is involved in several collagen-related disorders, including idiopathic pulmonary fibrosis.	(29,30)
SNTB1	Syntrophin, beta 1; dystrophin-associated protein A1, 59 kDa, basic component 1	Member of the syntrophin gene family, ubiquitously expressed and associated with dystrophin and related proteins.	(31)
SNX10	Sorting nexin 10	Modulates osteoclast differentiation; germ line mutations in this gene cause autosomal recessive osteopetrosis in humans.	(32,33)
TXNDC12	Thioredoxin domain containing 12	Member of the thioredoxin superfamily, whose other members are overexpressed in cancer or associated with tumor invasion.	(34,35)

Data are from the GeneCard database and other sources.

suggest the presence of a transcriptional profile, or signature, associated with overall survival in non-involved lung tissue of lung adenocarcinoma patients. Further studies with larger series of lung cancer patients are needed to establish whether or not a transcriptional signature of non-involved lung is predictive of the risk of poor survival in these patients. If such a hypothesis is verified, the prognosis of surgically treated lung cancer patients could be improved by a closer follow-up of those at higher risk of poor outcome. Hopefully, the identification of the involved genes and pathways will offer new therapeutic and prevention targets to improve these patients' survival.

Supplementary material

Supplementary Figures 1 and 2 can be found at http://carcin.oxfordjournals.org/

Funding

Italian Association for Cancer Research (AIRC; 10323 to T.A.D., 12162 to T.A.D. and Silvana Canevari).

Acknowledgements

We thank Dr Valerie Matarese for scientific editing and the entire staff of the core facility of Functional Genomics and Bioinformatics, Fondazione IRCCS Istituto Nazionale dei Tumori. In particular, we thank Dr Silvana Canevari, head of the facility, for support in the functional genomic analyses. The funders had no role in the design and conduct of the study; in the collection, analysis and interpretation of the data and in the preparation, review or approval of the manuscript.

Conflict of Interest Statement: None declared.

References

1. Goldstraw,P. *et al.* (2011) Non-small-cell lung cancer. *Lancet*, **378**, 1727–1740.
2. Marin,J.J. *et al.* (2012) Genetic variants in genes involved in mechanisms of chemoresistance to anticancer drugs. *Curr. Cancer Drug Targets*, **12**, 402–438.
3. Cadranel,J. *et al.* (2012) Impact of systematic EGFR and KRAS mutation evaluation on progression-free survival and overall survival in patients with advanced non-small-cell lung cancer treated by erlotinib in a French prospective cohort (ERMETIC project–part 2). *J. Thorac. Oncol.*, **7**, 1490–1502.
4. Johnson,M.L. *et al.* (2013) Association of KRAS and EGFR mutations with survival in patients with advanced lung adenocarcinomas. *Cancer*, **119**, 356–362.
5. Huang,Y.T. *et al.* (2009) Genome-wide analysis of survival in early-stage non-small-cell lung cancer. *J. Clin. Oncol.*, **27**, 2660–2667.
6. Lee,Y. *et al.* (2013) Prognostic implications of genetic variants in advanced non-small cell lung cancer: a genome-wide association study. *Carcinogenesis*, **34**, 307–313.
7. Franke,L. *et al.* (2009) eQTL analysis in humans. *Methods Mol. Biol.*, **573**, 311–328.
8. Kadara,H. *et al.* (2012) Pulmonary adenocarcinoma: a renewed entity in 2011. *Respirology*, **17**, 50–65.
9. Seo,J.S. *et al.* (2012) The transcriptional landscape and mutational profile of lung adenocarcinoma. *Genome Res.*, **22**, 2109–2119.
10. Park,Y.Y. *et al.* (2012) Development and validation of a prognostic gene-expression signature for lung adenocarcinoma. *PLoS One*, **7**, e44225.
11. Frullanti,E. *et al.* (2011) Multiple genetic loci modulate lung adenocarcinoma clinical staging. *Clin. Cancer Res.*, **17**, 2410–2416.
12. Frullanti,E. *et al.* (2012) Association of lung adenocarcinoma clinical stage with gene expression pattern in noninvolved lung tissue. *Int. J. Cancer*, **131**, E643–648.
13. Dassano,A. *et al.* (2013) Multigenic nature of the mouse pulmonary adenoma progression 1 locus. *BMC Genomics*, **14**, 152.
14. Spira,A. *et al.* (2004) Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proc. Natl Acad. Sci. USA*, **101**, 10143–10148.
15. Du,P. *et al.* (2008) lumi: a pipeline for processing Illumina microarray. *Bioinformatics*, **24**, 1547–1548.
16. Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
17. Trapnell,C. *et al.* (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
18. Robinson,J.T. *et al.* (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.
19. Zeggini,E. *et al.*; Wellcome Trust Case Control Consortium. (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.*, **40**, 638–645.
20. Normand,S.L. (1999) Meta-analysis: formulating, evaluating, combining, and reporting. *Stat. Med.*, **18**, 321–359.
21. Kadara,H. *et al.* (2012) Field cancerization in non-small cell lung cancer: implications in disease pathogenesis. *Proc. Am. Thorac. Soc.*, **9**, 38–42.
22. Goldie,S.J. *et al.* (2012) FRMD4A upregulation in human squamous cell carcinoma promotes tumor growth and metastasis and is associated with poor prognosis. *Cancer Res.*, **72**, 3424–3436.
23. Wang,T. *et al.* (2012) FERM-containing protein FRMD5 is a p120-catenin interacting protein that regulates tumor progression. *FEBS Lett.*, **586**, 3044–3050.
24. Wu,Y.M. *et al.* (2011) Mucin glycosylating enzyme GALNT2 regulates the malignant character of hepatocellular carcinoma by modifying the EGF receptor. *Cancer Res.*, **71**, 7270–7279.
25. Longobardi,E. *et al.* (2010) Prep1 (pKnox1)-deficiency leads to spontaneous tumor development in mice and accelerates EmuMyc lymphomagenesis: a tumor suppressor role for Prep1. *Mol. Oncol.*, **4**, 126–134.
26. Iotti,G. *et al.* (2011) Homeodomain transcription factor and tumor suppressor Prep1 is required to maintain genomic stability. *Proc. Natl Acad. Sci. USA*, **108**, E314–E322.
27. Schaeffer,P.J. *et al.* (2009) Impaired contractile function and calcium handling in hearts of cardiac-specific calcineurin b1-deficient mice. *Am. J. Physiol. Heart Circ. Physiol.*, **297**, H1263–H1273.
28. Goodyer,W.R. *et al.* (2012) Neonatal β cell development in mice and humans is regulated by calcineurin/NFAT. *Dev. Cell*, **23**, 21–34.
29. Ishida,Y. *et al.* (2011) Hsp47 as a collagen-specific molecular chaperone. *Methods Enzymol.*, **499**, 167–182.
30. Kakugawa,T. *et al.* (2013) Serum heat shock protein 47 levels are elevated in acute exacerbation of idiopathic pulmonary fibrosis. *Cell Stress Chaperones*, **18**, 581–590.
31. Johnson,E.K. *et al.* (2012) Proteomic analysis reveals new cardiac-specific dystrophin-associated proteins. *PLoS One*, **7**, e43515.
32. Zhu,C.H. *et al.* (2012) SNX10 is required for osteoclast formation and resorption activity. *J. Cell. Biochem.*, **113**, 1608–1615.
33. Pangrazio,A. *et al.* (2013) SNX10 mutations define a subgroup of human autosomal recessive osteopetrosis with variable clinical severity. *J. Bone Miner. Res.*, **28**, 1041–1049.
34. Vincent,E.E. *et al.* (2011) Overexpression of the TXNDC5 protein in non-small cell lung carcinoma. *Anticancer Res.*, **31**, 1577–1582.
35. Lu,A. *et al.* (2012) TXNDC9 expression in colorectal cancer cells and its influence on colorectal cancer prognosis. *Cancer Invest.*, **30**, 721–726.
36. Beer,D.G. *et al.* (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.*, **8**, 816–824.
37. Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma, *et al.* (2008) Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat. Med.*, **14**, 822–827.
38. Tomida,S. *et al.* (2009) Relapse-related molecular signature in lung adenocarcinomas identifies patients with dismal prognosis. *J. Clin. Oncol.*, **27**, 2793–2799.
39. Lu,Y. *et al.* (2012) Gene-expression signature predicts postoperative recurrence in stage I non-small cell lung cancer patients. *PLoS One*, **7**, e30880.

Received May 22, 2013; revised August 7, 2013; accepted August 18, 2013