33 Yousem SA, Ohori NP, Sonmez-Alpan E. Occurrence of human papillomavirus DNA in primary lung neoplasms. *Cancer* 1992; **69**: 693–7.
34 Wistuba II, Behrens C, Milchgrub S *et al.* Comparison of molecular changes in lung cancers in HIV-positive and HIV-indeterminate subjects. *JAMA* 1998; **279**: 1554–9.
35 Shimizu E, Coxon A, Otterson GA *et al.* RB protein status and clinical correlation from 171 cell lines representing lung cancer, extrapulmonary small cell carcinoma, and mesothelioma. *Oncogene* 1994; **9**: 2441–8.
36 Cheng YW, Wu MF, Wang J *et al.* Human papillomavirus 16/18 E6 oncoprotein is expressed in lung cancer and related with p53 inactivation. *Cancer Res* 2007; **67**: 10686–93.
37 Wang Y, Wang A, Jiang R *et al.* Human papillomavirus type 16 and 18 infection is associated with lung cancer patients from the central part of China. *Oncol Rep* 2008; **2**: 333–9.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Prevalence of human papillomavirus (HPV) 16, 18, and 33 in lung adenocarcinomas in East Asia.

Table S2. Primer sequences for detection of human papillomavirus (HPV) DNA in cancer cell DNA.

Table S3-1. p53 status defined in our studies but not registered in the COSMIC database.

Table S3-2. Concordance of p53 status defined in our studies and registered in the COSMIC database.

Table S3-3. Discordance of p53 status defined in our studies and registered in the COSMIC database.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

doi: 10.1111/j.1349-7006.2010.01622.x

Published OnlineFirst November 11, 2011; DOI:10.1158/0008-5472.CAN-11-1403

Molecular and Cellular Pathobiology

Cancer Research

Identification of Genes Upregulated in *ALK*-Positive and *EGFR/KRAS/ALK*-Negative Lung Adenocarcinomas

Hirokazu Okayama[1,2,10], Takashi Kohno[2], Yuko Ishii[1,2], Yoko Shimada[2], Kouya Shiraishi[2], Reika Iwakawa[1], Koh Furuta[5], Koji Tsuta[5], Tatsuhiro Shibata[3], Seiichiro Yamamoto[7], Shun-ichi Watanabe[6], Hiromi Sakamoto[4], Kensuke Kumamoto[10], Seiichi Takenoshita[10], Noriko Gotoh[8], Hideaki Mizuno[11,12], Akinori Sarai[11], Shuichi Kawano[9], Rui Yamaguchi[9], Satoru Miyano[9], and Jun Yokota[1]

Abstract

Activation of the *EGFR, KRAS,* and *ALK* oncogenes defines 3 different pathways of molecular pathogenesis in lung adenocarcinoma. However, many tumors lack activation of any pathway (triple-negative lung adenocarcinomas) posing a challenge for prognosis and treatment. Here, we report an extensive genome-wide expression profiling of 226 primary human stage I–II lung adenocarcinomas that elucidates molecular characteristics of tumors that harbor *ALK* mutations or that lack *EGFR, KRAS,* and *ALK* mutations, that is, triple-negative adenocarcinomas. One hundred and seventy-four genes were selected as being upregulated specifically in 79 lung adenocarcinomas without *EGFR* and *KRAS* mutations. Unsupervised clustering using a 174-gene signature, including *ALK* itself, classified these 2 groups of tumors into *ALK*-positive cases and 2 distinct groups of triple-negative cases (groups A and B). Notably, group A triple-negative cases had a worse prognosis for relapse and death, compared with cases with *EGFR, KRAS,* or *ALK* mutations or group B triple-negative cases. In *ALK*-positive tumors, 30 genes, including *ALK* and *GRIN2A,* were commonly overexpressed, whereas in group A triple-negative cases, 9 genes were commonly overexpressed, including a candidate diagnostic/therapeutic target *DEPDC1,* that were determined to be critical for predicting a worse prognosis. Our findings are important because they provide a molecular basis of *ALK*-positive lung adenocarcinomas and triple-negative lung adenocarcinomas and further stratify more or less aggressive subgroups of triple-negative lung ADC, possibly helping identify patients who may gain the most benefit from adjuvant chemotherapy after surgical resection. *Cancer Res; 72(1); 100–11. ©2011 AACR.*

Introduction

Lung cancer is the leading cause of cancer death worldwide (1, 2). Adenocarcinoma, which accounts for more than 50% of non-small-cell lung cancers (NSCLC), is the most frequent type and is increasing. Lung adenocarcinoma has a heterogeneous nature in various aspects, including clinicopathologic features (3). Recent molecular studies have revealed at least 3 major molecular pathways for the development of lung adenocarcinoma (4–8). A considerable fraction (30%–60%) of lung adenocarcinomas develops through acquisition of mutations either in the *EGFR, KRAS,* or *ALK* genes in a mutually exclusive manner, and the remaining lung adenocarcinomas, that is, those without *EGFR, KRAS,* and *ALK* mutations (herein designated "triple-negative adenocarcinomas"), develop with mutations of several other genes. *HER2, BRAF,* etc. are known to be mutated also mutually exclusively with the *EGFR, KRAS,* and *ALK* genes; however, frequencies of their mutations are very low (<5%; refs. 4–7). Therefore, genes responsible for the development of triple-negative adenocarcinomas are largely unknown.

Mutations in the *EGFR* gene are prevalent in females and never-smokers, and the frequencies are considerably higher in Asians (40%–60%) than in Europeans/Americans (~10%; refs. 5–7, 9). *EGFR* mutations make tumor cells dependent on epidermal growth factor receptor (EGFR) signaling and define patients who respond to EGFR tyrosine kinase inhibitors (TKI), such as gefitinib (10, 11). On the other hand, mutations in the *KRAS* gene occur predominantly in males and ever-smokers, and their frequencies are higher in Europeans/Americans (>15%) than in Asians (10%; ref. 9). Specific inhibitors against KRAS activity are being developed (12). Therefore, clinicopathologic features of lung adenocarcinomas with *EGFR* mutations (herein designated "*EGFR*-positive adenocarcinomas") and

Authors' Affiliations: Divisions of [1]Multistep Carcinogenesis, [2]Genome Biology, [3]Cancer Genomics and [4]Genetics, National Cancer Center Research Institute; [5]Division of Pathology and Clinical Laboratories and [6]Thoracic Surgery Division, National Cancer Center Hospital; [7]Cancer Information Services and Surveillance Division, Center for Cancer Control and Information Services, National Cancer Center; [8]Division of Systems Biomedical Technology and [9]Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo; [10]Department of Organ Regulatory Surgery, Fukushima Medical University School of Medicine, Fukushima; [11]Department of Biosciences and Bioinformatics, Kyushu Institute of Technology, Fukuoka; and [12]Discovery Science & Technology Department, Chugai Pharmaceutical Co., Ltd., Kanagawa, Japan

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Corresponding Author: Jun Yokota, Divison of Multistep Carcinogenesis, National Cancer Center Research Institute, Tsukiji 5-1-1, Chuo-ku, Tokyo 104-0045, Japan. Phone: 81-3-3547-5272; Fax: 81-3-3542-0807; E-mail: jyokota@ncc.go.jp

doi: 10.1158/0008-5472.CAN-11-1403

Published OnlineFirst November 11, 2011; DOI:10.1158/0008-5472.CAN-11-1403

Cancer Research

Identification of Genes Upregulated in *ALK*-Positive and *EGFR/KRAS/ALK*-Negative Lung Adenocarcinomas

Hirokazu Okayama, Takashi Kohno, Yuko Ishii, et al.

Cancer Res 2012;72:100-111. Published OnlineFirst November 11, 2011.

Published OnlineFirst November 11, 2011; DOI:10.1158/0008-5472.CAN-11-1403

those with *KRAS* mutations (herein designated "*KRAS*-positive adenocarcinomas") are considerably different from each other. Recently, a small subset of *EGFR*- and *KRAS*-negative lung adenocarcinomas (~5%) was shown to have rearrangements of the *ALK* gene generating gene fusion transcripts (13), and patients with *ALK* rearrangements tend to be younger and have little or no smoking histories (4, 6–8). Because lung adenocarcinoma cells with *ALK* rearrangements (herein designated "*ALK*-positive adenocarcinomas") are specifically sensitive to ALK TKIs, *ALK*-positive adenocarcinomas have been recently considered to be another subset of adenocarcinomas by considering the differences in therapeutic targets (4, 6–8). In contrast, clinicopathologic features of triple-negative lung adenocarcinomas have not been precisely characterized because of the lack of sufficient genetic information in these adenocarcinomas.

There have been several studies which attempted to characterize gene expression profiles in particular types of lung adenocarcinoma, including *EGFR*-positive and *KRAS*-positive adenocarcinomas (14–17). However, such information is limited for *ALK*-positive adenocarcinomas and triple-negative adenocarcinomas. Therefore, in this study, we aimed to elucidate clinicopathologic features and gene expression profiles of *ALK*-positive adenocarcinomas and triple-negative adenocarcinomas in comparison with those of *EGFR*-positive adenocarcinomas and *KRAS*-positive adenocarcinomas. We conducted a genome-wide gene expression profiling of 226 lung adenocarcinomas, consisting of 127 *EGFR*-positive adenocarcinomas, 20 *KRAS*-positive adenocarcinomas, 11 *ALK*-positive adenocarcinomas, and 68 triple-negative adenocarcinomas. To identify genes useful for molecular diagnosis and applicable to targeted therapy of *ALK*-positive adenocarcinomas and triple-negative adenocarcinomas, we focused on genes that were upregulated in these adenocarcinomas by selecting genes with low expression in *EGFR*-positive and *KRAS*-positive adenocarcinomas. Several genes were identified as being specifically and significantly upregulated in *ALK*-positive adenocarcinomas. In particular, the *ALK* gene itself was highly expressed exclusively in *ALK*-positive adenocarcinomas. More importantly, a distinct group of triple-negative adenocarcinomas with unfavorable outcome was identified. This group of triple-negative adenocarcinomas showed much worse prognosis than the other group of triple-negative adenocarcinomas, *EGFR*-positive adenocarcinomas, *KRAS*-positive adenocarcinomas, and *ALK*-positive adenocarcinomas. Several genes were identified as being upregulated and critical for predicting prognosis of patients in this group of adenocarcinomas.

Materials and Methods

Patients

The tumors were pathologically classified according to the TNM classification of malignant tumors (18). A total of 226 lung adenocarcinoma cases subjected to expression profiling were selected from 393 stage I–II cases who underwent potential curative resection between 1998 and 2008 at the National Cancer Center Hospital as follows (ref. 19; Supplementary Fig. S1). Among the 393 cases, 363 cases, consisting of 305 stage I and 58 stage II cases, were eligible by the criteria of cases who did not receive any neoadjuvant therapies before surgery and had not been diagnosed with cancer in the 5 years before lung adenocarcinoma diagnosis. All 58 stage II cases were subjected to expression profiling. The 305 stage I cases included 37 cases with relapse and 268 cases without relapse. To improve statistical efficiency, all the 37 relapsed cases and 131 matched unrelapsed cases selected by the incidence density sampling method (20, 21) were subjected to expression profiling. In total, 226 cases, consisting of 168 stage I and 58 stage II cases, were subjected to the expression profiling. Among the 226 cases, 204 who received complete resection (i.e., free resection margins and no involvement of mediastinal lymph nodes examined by mediastinal dissection) and did not receive postoperative chemotherapy and/or radiotherapy, unless relapsed, were subjected to survival analyses. This study was approved by the Institutional Review Boards of the National Cancer Center.

Microarray experiments and data processing

Total RNA was extracted using TRIzol reagent (Invitrogen), purified by an RNeasy kit (Qiagen), and qualified with a model 2100 Bioanalyzer (Agilent). All samples showed RNA Integrity Numbers more than 6.0 and were subjected to microarray experiments. Two micrograms of total RNA were labeled using a 5X MEGAscript T7 Kit (Ambion) and analyzed by Affymetrix U133Plus2.0 arrays. The data were processed by the MAS5 algorithm, and the mean expression level of a total of 54,675 probes was adjusted to 1,000 for each sample. Microarray data are available at National Center for Biotechnology Information Gene Expression Omnibus (GSE31210).

Probe selection for unsupervised clustering

One hundred and seventy-four genes (190 probes), preferentially expressed in *ALK*-positive and triple-negative adenocarcinomas, were selected by the following criteria; probes whose expression levels were less than 1,000 in any adenocarcinomas with *EGFR* or *KRAS* mutations, and probes whose averaged expression levels in *ALK*-positive and triple-negative adenocarcinomas were more than 1.5-fold higher than those in *EGFR*-positive and *KRAS*-positive adenocarcinomas with P values less than 0.05 by t test. Expression levels for these 190 probes were log-transformed and median-centered, both for probes and samples, and were subjected to an unsupervised hierarchical clustering. The clustering was done by the centroid linkage method using the Cluster 3.0 program, and the results were visualized using the Java Treeview program (22).

Mutation analyses

Genomic DNAs from all 226 lung adenocarcinomas were analyzed for *EGFR* and *KRAS* mutations by the high-resolution melting method as described (23, 24). Total RNAs from the 226 adenocarcinomas were examined for expression of fusion transcripts between *ALK* and *EML4* or *KIF5* using a multiplex reverse transcription PCR (RT-PCR) method (25).

Statistics

Cumulative survival was estimated by the Kaplan–Meier method, and differences in the survivals between 2 groups were

Published OnlineFirst November 11, 2011; DOI:10.1158/0008-5472.CAN-11-1403

analyzed by log-rank test. Influences of variables on relapse-free survival (RFS) and overall survival (OS) were evaluated by uni- and multivariate analyses of the Cox proportional hazard model. For all analyses, smoking status was polarized as never-smokers (0 pack years) and ever-smokers (>0 pack years). Pathologic TNM staging was categorized as stage I versus stage II. For multivariate analysis, all variables were included that were moderately associated ($P < 0.1$) with RFS or OS in any of the analyses.

Bioinformatics

Associations of gene expression levels with prognosis of NSCLC patients in 7 other expression profile studies were obtained from the PrognoScan database (26). In the PrognoScan database, association of gene expression with survival of patients was evaluated by the minimum P value approach. Briefly, patients were first arranged by expression levels of a given gene. They were then divided into high- and low-expression groups at all possible cutoff points, and the risk differences of any 2 groups were estimated by the log-rank test. Finally, the cutoff point that gave the most pronounced P value was selected.

Results

EGFR/KRAS/ALK mutations and clinicopathologic characteristics of lung adenocarcinomas subjected to gene expression profiling

Among 226 stages I and II lung adenocarcinomas, *EGFR* and *KRAS* mutations were mutually exclusively detected in 127 (56%) and 20 (9%) cases, respectively, and an *EML4–ALK* fusion gene was expressed in 11 (4.9%) cases (Table 1). *EGFR* or *KRAS* mutations were not detected in any of the 11 cases with *EML4–ALK* fusion expression; thus, the occurrence of *ALK* rearrangements in a mutually exclusive manner with *EGFR* and *KRAS* mutations in lung adenocarcinoma was confirmed. The incidence and the fraction of *EGFR*-, *KRAS*-, and *ALK*-positive cases in this study were consistent with those in previous studies (5–7, 9, 13). Accordingly, the remaining 68 (30%) cases were defined as "triple-negative adenocarcinomas" because of the absence of *EGFR*, *KRAS*, and *ALK* mutations. Clinicopathologic features of *EGFR*-positive adenocarcinomas and *KRAS*-positive adenocarcinomas in this study are well consistent with those in previous studies of Japanese populations (27, 28). Patients with *ALK*-positive adenocarcinomas were younger and more likely to be never-smokers, as previously indicated (4, 6–8). Triple-negative adenocarcinomas showed similar clinicopathologic features to those of *KRAS*-positive adenocarcinomas, that is, a predominance of males, ever-smokers, and advanced stages.

Expression profile unique to *ALK*-positive lung adenocarcinomas

All 226 cases were subjected to genome-wide expression profiling using Affymetrix U133Plus2.0 arrays. One hundred and seventy-four genes, evaluated with 190 probes (Supplementary Table S1), were selected as those preferentially expressed in either *ALK*-positive adenocarcinomas or triple-negative adenocarcinomas under the criteria described in Materials and Methods. In particular, 10 genes evaluated with 11 probes were markedly upregulated according to the criteria of fold-differences more than 2.0 with P values less than 0.05 (Supplementary Table S2). It was noted that 2 probes for the *ALK* gene were present among them, and 1 of them (probe ID = 208212_s_at) showed the highest fold-difference of 8.7 between *ALK*-positive/triple-negative adenocarcinomas and *EGFR*-positive/*KRAS*-positive adenocarcinomas among the 190 probes. This result indicated that there is a subset of adenocarcinomas in which *ALK* was overexpressed. Therefore, an unsupervised

Table 1. Clinicopathologic characteristics of 226 lung adenocarcinomas subjected to expression profile analysis

		Mutation				Expression profile	
Variable	All	EGFR (+)	KRAS (+)	ALK (+)	Triple (−)	Group A	Group B
No. of cases	226	127	20	11	68	36	32
Age							
Mean	60	60	60	54	61	61	60
Range	30–76	35–72	46–75	30–68	46–76	46–71	47–76
Sex							
Male	105	50	12	2	41	25	16
Female	121	77	8	9	27	11	16
Smoking habit							
Never-smoker	115	67	10	7	31	10	21
Ever-smoker	111	60	10	4	37	26	11
pStage							
IA	114	77	6	3	28	10	18
IB	54	26	8	0	20	12	8
II	58	24	6	8	20	14	6

Published OnlineFirst November 11, 2011; DOI:10.1158/0008-5472.CAN-11-1403

Table 2. Genes upregulated in *ALK*-positive lung adenocarcinomas

Gene symbol[a]	Gene name	Probe ID	Fold difference
ALK	Anaplastic lymphoma receptor tyrosine kinase	208212_s_at	55.2
EST	Transcribed locus	242964_at	26.8
ALK	Anaplastic lymphoma receptor tyrosine kinase	208211_s_at	17.2
GRIN2A	Glutamate receptor, ionotropic, *N*-methyl D-aspartate 2A	242286_at	13.0
GRIN2A	Glutamate receptor, ionotropic, *N*-methyl D-aspartate 2A	231384_at	12.4
MUC5AC /// MUC5B	Mucin 5AC, oligomeric mucus/gel-forming /// mucin 5B, oligomeric mucus/gel-forming	222268_x_at	9.2
EST	Transcribed locus	1570291_at	8.1
LOC100292909	Hypothetical protein LOC100292909	241535_at	7.7
BLID	BH3-like motif containing, cell death inducer	1555675_at	7.4
LOC100130894	Hypothetical LOC100130894	1564158_a_at	6.1
CLDN10	Claudin 10	1556687_a_at	6.0
KRT16	Keratin 16	209800_at	5.9
PROM2	Prominin 2	1562378_s_at	5.6
GJB5	Gap junction protein, beta 5, 31.1 kDa	206156_at	5.0
KIAA1644	KIAA1644	221901_at	4.8
EPHB1	EPH receptor B1	210753_s_at	4.5
LRRC4	Leucine rich repeat containing 4	223552_at	4.2
EST	Transcribed locus	235373_at	3.4
tcag7.1188	Hypothetical LOC340340	1561254_at	3.3
SBNO2	Strawberry notch homolog 2 (*Drosophila*)	204166_at	3.3
EST	Transcribed locus	241083_at	3.1
SLC25A37	Solute carrier family 25, member 37	222528_s_at	3.1
NDP	Norrie disease (pseudoglioma)	206022_at	3.1
EST	Transcribed locus	243478_at	3.0
EST	Transcribed locus	239136_at	2.9
RHOV	ras homolog gene family, member V	241990_at	2.9
YIF1B	Yip1 interacting factor homolog B (*S. cerevisiae*)	231211_s_at	2.9
RPRM	Reprimo, TP53 dependent G2 arrest mediator candidate	219370_at	2.5
SYT12	Synaptotagmin XII	228072_at	2.5
HES2	Hairy and enhancer of split 2 (*Drosophila*)	231928_at	2.4
CDH11	Cadherin 11, type 2, OB-cadherin (osteoblast)	239769_at	2.2
IRAK3	Interleukin-1 receptor-associated kinase 3	220034_at	2.1

[a]Genes with fold difference >2.0 and $P < 0.05$ between *ALK*-positive and *ALK*-negative adenocarcinomas are shown.

hierarchical clustering using these 190 probes was done on 11 *ALK*-positive adenocarcinomas and 68 triple-negative adenocarcinomas (Supplementary Figs. S1 and S2). There were 3 distinct sets of genes/probes, as indicated by red, yellow, and blue bars on the left of the heat map. Two probes for the *ALK* gene were present in the gene/probe set with a yellow bar, and 11 cases with extremely high levels of *ALK* expression comprised a small subcluster in the right side of cluster 1. All the 11 cases corresponded to the ones with *EML4–ALK* fusion gene expression.

The results strongly indicated that *ALK*-positive adenocarcinomas have distinct expression profiles in comparison with *ALK*-negative adenocarcinomas, including not only triple-negative adenocarcinomas but also *EGFR*-positive and *KRAS*-positive adenocarcinomas. Therefore, genes with fold-differences more than 2.0 and P values less than 0.05 in their expression between *ALK*-positive adenocarcinomas and *ALK*-negative adenocarcinomas were further selected from the 190 probes. Thirty genes with 32 probes were then selected (Table 2). The *ALK* gene showed the highest level of fold difference in *ALK*-positive adenocarcinomas. Therefore, as previously reported (29–31), *ALK*-positive adenocarcinomas express high levels of *ALK* gene products, supporting that upregulation of the *ALK* gene is a biological consequence of *ALK* rearrangements in lung adenocarcinoma cells. Expression profiling further revealed that various other genes are distinctly upregulated in *ALK*-positive adenocarcinomas. In particular, fold differences of *GRIN2A* (glutamate receptor, ionotropic, *N*-methyl D-aspartate 2A) expression were more than 10, as with *ALK* expression. Moreover, *GRIN2A* was branched most closely to *ALK* in the heat map (Supplementary Fig. S2). Therefore, high levels of *GRIN2A* expression can be a characteristic unique to *ALK*-positive adenocarcinomas, in addition to upregulation of the *ALK* gene itself. The levels of *GRIN2A* expression in *ALK*-

Published OnlineFirst November 11, 2011; DOI:10.1158/0008-5472.CAN-11-1403

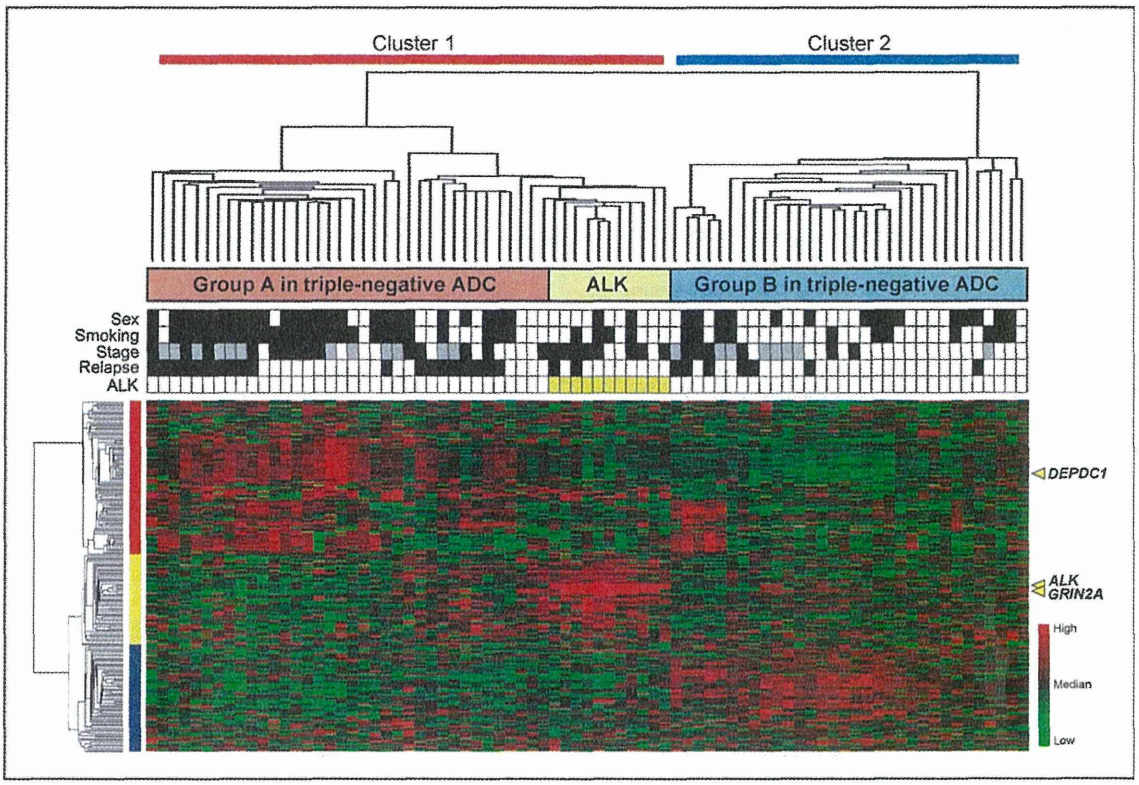

Figure 1. Unsupervised hierarchical clustering of 11 *ALK*-positive adenocarcinomas and 68 triple-negative adenocarcinomas. Triple-negative adenocarcinomas were separated into 36 group A cases and 32 group B cases, and group A cases construct cluster 1 with 11 *ALK*-positive adenocarcinoma cases. Clinical and genetic features are shown below the tree; sex (black, male; white, female); smoking status (black, ever-smoker; white, never-smoker); pathologic stage (black, stage II; gray, stage IB; white, stage IA); relapse (black, evidence of relapse; white, no evidence of relapse); *ALK* (yellow, *ALK*-fusion gene expression positive; white, negative). Three colored bars according to the main branches of probes/genes are shown on the left. Positions of probes for *ALK*, *GRIN2A*, and *DEPDC1* are shown on the right. ADC, adenocarcinoma.

positive adenocarcinomas were significantly higher than those in *ALK*-negative adenocarcinomas by quantitative RT-PCR analysis (Supplementary Fig. S3).

Triple-negative lung adenocarcinomas with poor prognosis identified by gene expression profiling

By the unsupervised hierarchical clustering, 68 triple-negative adenocarcinomas were separated into 2 major groups, one containing 36 cases and the other 32 cases, designated as groups A and B, respectively (Fig. 1). Group A comprised cluster 1 with 11 *ALK*-positive adenocarcinomas. Group A cases were dominant in males, ever-smokers, and advanced stages, whereas group B cases were dominant in never-smokers and early stages (Table 1), indicating that group A cases comprise an aggressive type in triple-negative adenocarcinomas. Therefore, we next compared RFS and OS among the 5 groups of patients; groups A and B, *EGFR*-positive cases, *KRAS*-positive cases, and *ALK*-positive cases (Fig. 2). Among the 226 cases, 204 cases that received complete resection and did not receive postoperative chemotherapy and/or radiotherapy were subjected to survival analysis. Group A cases ($n = 32$) showed the worst prognosis for both RFS and OS among the 5 groups (Fig. 2A and B). In particular, group A cases showed significantly worse prognosis ($P < 0.05$) for both RFS and OS than group B cases ($n = 30$) and *EGFR*-positive cases ($n = 116$) by the log-rank test. Such differences were marginally significant between group A cases and *KRAS*-positive cases ($n = 19$) and not significant between group A cases and *ALK*-positive cases ($n = 7$), probably because the numbers of *KRAS*-positive and *ALK*-positive cases were smaller than those of group B and *EGFR*-positive cases.

Similar results were obtained from the analysis of 162 patients with stage I adenocarcinomas (Fig. 2C and D), indicating the independency of these associations with staging. Therefore, we next carried out multivariate analyses on RFS and OS of these 5 groups (Table 3). In the analysis of 204 stages I and II patients, RFS and OS of group A cases were significantly worse than those of *EGFR*-positive and group B cases, and the differences were independent of staging. HRs of *ALK*-positive and *KRAS*-positive cases were also as high as *EGFR*-positive and group B cases, although only the difference in RFS was statistically significant between group A cases and *KRAS*-positive cases. This could be also due to the small numbers

Published OnlineFirst November 11, 2011; DOI:10.1158/0008-5472.CAN-11-1403

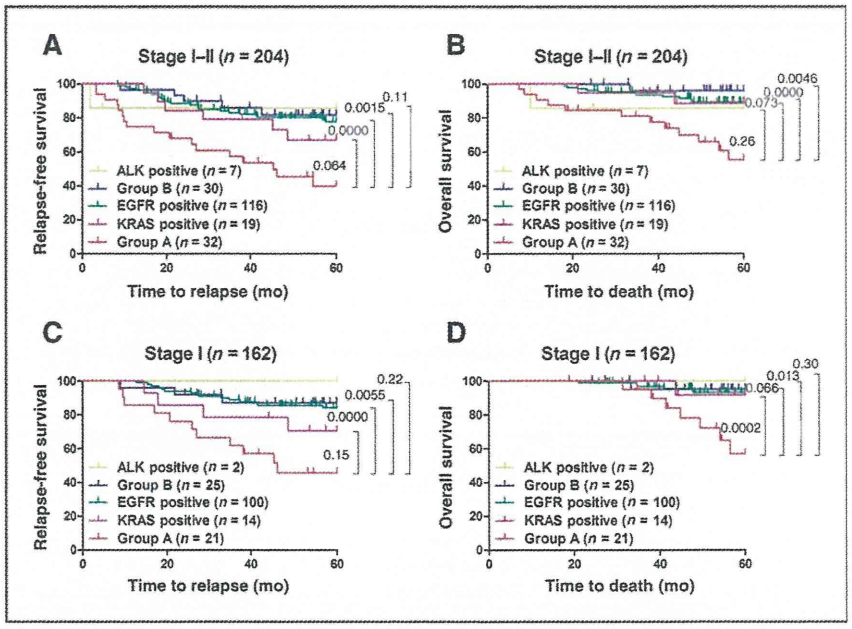

Figure 2. Kaplan–Meier survival curves for RFS and OS of 204 lung adenocarcinoma cases according to *EGFR*-positive, *KRAS*-positive, *ALK*-positive, group A, and group B. RFS and OS of stage I–II (A, B) and stage I (C, D) cases are shown.

of *KRAS*-positive and *ALK*-positive cases. Accordingly, multivariate analyses of 162 stage I patients further showed significant differences in RFS and OS between group A cases and *EGFR*-positive cases, and also between group A cases and group B cases. Because numbers of *KRAS*-positive cases and *ALK*-positive cases were small, we next compared RFS and OS between group A patients and patients in all 4 other groups combined ("Others" in Table 3). Differences in RFS as well as those in OS were highly significant and independent of staging. These results strongly indicated that group A patients comprise a distinct subclass of *EGFR/KRAS/ALK*-negative lung adenocarcinomas, and the prognoses of group A patients were the worst among the 5 groups of patients.

Clustering of lung adenocarcinomas with poor prognosis by gene expression profiling

We next carried out unsupervised hierarchical clustering of all the 226 adenocarcinoma cases, including 127 *EGFR*-positive cases and 20 *KRAS*-positive cases, to investigate whether expression profiling with a set of 174 genes with 190 probes could extract group A cases as a unique subset among all adenocarcinomas, and whether the profiling could be useful for prognosis prediction of patients with any genotypes of adenocarcinomas in general. As shown in Supplementary Fig. S4, clustering patterns of all the 226 patients were very similar to those of the 79 patients consisting of 11 *ALK*-positive cases and 68 triple-negative cases. In particular, the 11 *ALK*-positive cases comprised a small cluster in the right side of Cluster 1 (Cluster 1b), supporting that *ALK*-positive adenocarcinomas show unique expression profiles among all adenocarcinomas. Group A and group B cases also have a tendency to accumulate in Clusters 1a and Cluster 2, respectively. However, group A cases often comprise clusters with the *KRAS*-positive cases, whereas group B cases were distributed with the *EGFR*-positive cases. Therefore, group A and group B triple-negative adenocarcinomas were not exclusive with the *EGFR*-positive and *KRAS*-positive adenocarcinomas by expression profiling of these 174 genes. Therefore, expression profiling with a set of the 174 genes was concluded to be useful to distinguish *ALK*-positive adenocarcinomas among all lung adenocarcinomas.

However, RFS of 119 patients in Cluster 1 was significantly worse than RFS of 85 patients in Cluster 2 (HR = 3.73, $P = 0.00016$). When Cluster 1 was further divided into 2 subclasses 1a and 1b of the right and left sides, respectively, Cluster 1a containing most of group A patients showed the worst prognosis among the 3 subclasses (Supplementary Fig. S4). Therefore, the expression signature of these 174 genes was indicated to be useful for prognostic prediction of adenocarcinoma patients, in particular of triple-negative adenocarcinoma patients.

Minimum set of genes characterizing triple-negative lung adenocarcinomas with poor prognosis

The above results implied that triple-negative adenocarcinomas can be classified into 2 distinct subgroups by expression profiling and prognoses of these 2 groups are significantly different from each other. Accordingly, expression of several genes among the 174 genes was expected to be independently associated with prognosis of triple-negative adenocarcinoma patients. Therefore, we next selected genes whose expression was associated with prognosis from the 174 genes evaluated by the 190 probes. To evaluate the prognostic value of each probe and to make a comparative study for association of gene expression with prognosis in other cohorts possible, we took a minimum P value approach for grouping the patients for survival analysis because of the following reason. A database

Published OnlineFirst November 11, 2011; DOI:10.1158/0008-5472.CAN-11-1403

Table 3. Hazard ratios for relapse-free and overall survivals in lung adenocarcinomas

			Univariate		Multivariate	
Survival	**Case (*n*)**	**Variable**	**HR (95% CI)**	***P***	**HR (95% CI)**	***P***
Relapse free	Stage I–II (204)	Age	1.03 (0.99–1.07)	0.12	1.04 (0.99–1.08)	0.092
		Sex (male/female)	1.39 (0.82–2.38)	0.22	1.00 (0.49–2.05)	0.99
		Smoking habit (ever/never)	1.43 (0.84–2.44)	0.19	1.10 (0.54–2.24)	0.80
		pStage (II/I)	1.86 (1.41–2.45)	1.3E-05	3.50 (1.93–6.34)	3.6E-05
		Subgroup				
		Group A/ALK (+)	4.78 (0.63–35.99)	0.13	6.01 (0.76–47.82)	0.09
		Group A/KRAS (+)	2.43 (0.96–6.17)	0.062	2.85 (1.10–7.35)	0.031
		Group A/EGFR (+)	3.58 (1.93–6.64)	5.3E-05	2.76 (1.44–5.29)	0.0022
		Group A/Group B	4.58 (1.69–12.42)	0.0028	4.10 (1.50–11.24)	0.0061
		Group A/Others	3.56 (2.00–6.34)	1.6E-05	3.04 (1.68–5.53)	2.5E-04
	Stage I (162)	Age	1.01 (0.96–1.06)	0.69	1.00 (0.95–1.05)	0.97
		Sex (male/female)	0.99 (0.50–1.96)	0.98	0.83 (0.33–2.07)	0.69
		Smoking habit (ever/never)	1.06 (0.54–2.08)	0.87	0.97 (0.39–2.45)	0.95
		Subgroup				
		Group A/ALK (+)	—	—	—	—
		Group A/KRAS (+)	2.31 (0.73–7.28)	0.15	2.36 (0.73–7.62)	0.15
		Group A/EGFR (+)	4.33 (2.00–9.35)	2.0E-04	4.51 (2.05–9.91)	1.7E-04
		Group A/Group B	5.36 (1.49–19.24)	0.010	5.52 (1.50–20.37)	0.010
		Group A/Others	4.18 (2.03–8.60)	1.0E-04	4.32 (2.06–9.09)	1.1E-04
Overall	Stage I–II (204)	Age	1.03 (0.98–1.08)	0.33	1.03 (0.98–1.09)	0.21
		Sex (male/female)	1.69 (0.82–3.48)	0.16	0.89 (0.33–2.41)	0.82
		Smoking habit (ever/never)	1.91 (0.92–3.97)	0.084	1.46 (0.54–3.92)	0.45
		pStage (II/I)	2.07 (1.45–2.97)	7.2E-05	3.93 (1.83–8.44)	4.6E-04
		Subgroup				
		Group A/ALK (+)	2.95 (0.38–22.78)	0.30	3.50 (0.41–29.85)	0.25
		Group A/KRAS (+)	3.12 (0.88–11.09)	0.079	3.31 (0.91–12.03)	0.069
		Group A/EGFR (+)	4.59 (2.06–10.23)	2.0E-04	3.35 (1.44–7.81)	0.005
		Group A/Group B	6.83 (1.53–30.54)	0.012	5.68 (1.24–25.95)	0.025
		Group A/Others	4.50 (2.17–9.36)	5.7E-05	3.61 (1.68–7.78)	0.0010
	Stage I (162)	Age	0.99 (0.93–1.06)	0.73	0.98 (0.91–1.04)	0.45
		Sex (male/female)	1.15 (0.43–3.08)	0.79	0.77 (0.20–3.00)	0.70
		Smoking habit (ever/never)	1.47 (0.55–3.91)	0.45	1.26 (0.32–4.89)	0.74
		Subgroup				
		Group A/ALK (+)	—	—	—	—
		Group A/KRAS (+)	5.79 (0.71–47.3)	0.10	5.61 (0.67–46.84)	0.11
		Group A/EGFR (+)	5.83 (2.04–16.71)	0.0010	6.06 (2.08–17.71)	9.8E-04
		Group A/Group B	9.13 (1.12–74.34)	0.039	9.32 (1.10–78.61)	0.040
		Group A/Others	6.30 (2.34–16.99)	2.8E-04	6.47 (2.33–17.98)	3.4E-04

named PrognoScan was recently developed by coauthors of this study (26). In the PrognoScan database, minimum *P* values for the association of gene expression with prognosis of all probes in a platform are available for a number of cohorts that have been published. Therefore, it was possible to validate the present findings using data from various other cohorts by the same criteria. According to the method described previously (26), corrected minimum *P* values were calculated for each probe to control the error rate for the evaluation of the association with RFS and OS. Expression of 11 genes evaluated with 12 probes (2 probes for the *DEPDC1* gene) showed significant associations with both RFS and OS in 62 triple-negative adenocarcinomas and also in 46 stage I triple-negative adenocarcinomas (Table 4). Among the 11 genes, expression of 10 genes was positively correlated with poor prognosis, whereas that of the remaining 1 gene, *KIF19*, expression was negatively correlated with poor prognosis.

We first selected 174 genes as being preferentially expressed in either *ALK*-positive adenocarcinomas or triple-negative adenocarcinomas by the criteria of "probes whose expression levels in any adenocarcinomas with *EGFR* or *KRAS* mutations were lower than the mean expression

Table 4. List of genes whose expression is associated with relapse free survival and overall survival of patients with lung adenocarcinoma

		NCC														CAN/DF		HLM		MSK		UM		Nagoya		Duke		Seoul	
		TN, Stage I–II				TN, Stage I				All Stage I–II				All Stage I		Stage I–III		Stage I–III		Stage I–III		Stage I–III		Stage I–III		Stage I–III		Stage I–III	
Dataset		RFS $n = 62$		OS $n = 62$		RFS $n = 46$		OS $n = 46$		RFS $n = 204$		OS $n = 204$		RFS $n = 162$		OS $n = 162$		OS $n = 82$		OS $n = 79$		OS $n = 104$		OS $n = 178$		OS $n = 117$		RFS $n = 111$		RFS $n = 138$	
Gene symbol	Probe ID (for NCC)	*P*	HR	*P*	HR	*P*	HR	*P*	HR	*P*	HR	*P*	HR	*P*	HR	*P*	HR	*P*	HR	*P*	HR	*P*	HR	*P*	HR	*P*	HR	*P*	HR	*P*	HR
DEPDC1	222958_s_at	0.00	2.3	0.00	3.0	0.00	3.0	0.02	2.7	0.00	2.1	0.00	1.8	0.00	2.0	0.00	2.1	—	—	—	—	0.03	1.1	—	—	0.00	1.6	0.04	1.0	0.01	0.9
DEPDC1	235545_at	0.00	1.8	0.01	2.3	0.00	2.4	0.04	2.6	0.00	1.4	0.00	1.6	0.00	1.3	0.00	2.2														
FOSL2	218881_s_at	0.01	1.7	0.03	1.7	0.02	1.8	0.00	3.3	0.00	1.2	0.00	1.7	0.00	1.4	0.00	2.4	—	—	0.00	1.7	—	—	0.02	0.7	—	—	—	—	0.01	1.0
MCM4	222037_at	0.00	1.8	0.00	3.0	0.01	2.0	0.04	2.6	0.00	1.4	0.00	1.8	0.00	1.5	0.00	2.1	—	—	—	—	—	—	—	—	0.00	1.7	—	—	—	—
UBE2S	202779_s_at	0.00	3.0	0.02	16.0	0.01	2.8	0.02	16.6	0.00	1.6	0.02	1.4	0.00	1.6	0.00	2.1	—	—	—	—	0.05	1.0	—	—	—	—	—	—	—	—
CD300A	217078_s_at	0.01	1.7	0.00	2.1	0.04	1.7	0.00	2.8	0.00	1.1	0.00	1.5	0.01	1.2	0.01	1.7	—	—	—	—	0.00	1.5	—	—	—	—	—	—	—	—
SLITRK4	232636_at	0.02	1.7	0.03	1.7	0.00	2.9	0.00	2.5	0.01	1.1	0.00	1.4	0.04	1.1	0.00	2.0	—	—	—	—	—	—	—	—	—	—	—	—	—	—
KRT16	209800_at	0.00	2.0	0.00	2.5	0.00	2.4	0.00	2.7	0.00	1.2	0.00	1.4	0.01	1.2	0.01	1.7	—	—	—	—	—	—	—	—	—	—	—	—	—	—
SIGLEC9	210569_s_at	0.00	1.9	0.01	2.0	0.00	2.1	0.04	2.1	0.00	1.6	0.00	1.4	0.00	1.7	0.00	2.2	—	—	—	—	—	—	—	—	—	—	—	—	—	—
DIAPH3	232596_at	0.02	1.5	0.00	3.0	0.05	1.6	0.03	2.6	0.00	1.2	0.00	2.1	0.00	1.5	0.00	2.0	—	—	—	—	—	—	—	—	—	—	—	—	—	—
LOC152225	1562048_at	0.01	1.5	0.00	2.3	0.02	1.7	0.00	2.7	0.00	1.3	0.00	1.9	—	—	0.00	1.9	—	—	—	—	—	—	—	—	—	—	—	—	—	—
KIF19	1553314_a_at	0.01	-1.5	0.05	-1.6	0.00	-3.0	0.00	−2.5	—	—	0.03	−1.4	—	—	—	—	—	—	—	—	—	—	—	—	0.00	−1.4	—	—	—	—

Abbreviations: NCC, National Cancer Center; TN, Triple-negative.

HRs (log2 ratio) with corrected *P* value < 0.05 are shown.