*Imaging, Diagnosis, Prognosis*

Clinical Cancer Research

# Biomarker Discovery in Non–Small Cell Lung Cancer: Integrating Gene Expression Profiling, Meta-analysis, and Tissue Microarray Validation

Johan Botling[1], Karolina Edlund[1], Miriam Lohr[9], Birte Hellwig[9], Lars Holmberg[2,4,11], Mats Lambe[4,5], Anders Berglund[4,5], Simon Ekman[3], Michael Bergqvist[3], Fredrik Pontén[1], André König[9], Oswaldo Fernandes[6], Mats Karlsson[7], Gisela Helenius[7], Christina Karlsson[8], Jörg Rahnenführer[9], Jan G Hengstler[10], and Patrick Micke[1]

## Abstract

**Purpose:** Global gene expression profiling has been widely used in lung cancer research to identify clinically relevant molecular subtypes as well as to predict prognosis and therapy response. So far, the value of these multigene signatures in clinical practice is unclear, and the biologic importance of individual genes is difficult to assess, as the published signatures virtually do not overlap.

**Experimental Design:** Here, we describe a novel single institute cohort, including 196 non–small lung cancers (NSCLC) with clinical information and long-term follow-up. Gene expression array data were used as a training set to screen for single genes with prognostic impact. The top 450 probe sets identified using a univariate Cox regression model (significance level $P < 0.01$) were tested in a meta-analysis including five publicly available independent lung cancer cohorts ($n = 860$).

**Results:** The meta-analysis revealed 14 genes that were significantly associated with survival ($P < 0.001$) with a false discovery rate <1%. The prognostic impact of one of these genes, the cell adhesion molecule 1 (CADM1), was confirmed by use of immunohistochemistry on tissue microarrays from 2 independent NSCLC cohorts, altogether including 617 NSCLC samples. Low CADM1 protein expression was significantly associated with shorter survival, with particular influence in the adenocarcinoma patient subgroup.

**Conclusions:** Using a novel NSCLC cohort together with a meta-analysis validation approach, we have identified a set of single genes with independent prognostic impact. One of these genes, CADM1, was further established as an immunohistochemical marker with a potential application in clinical diagnostics. *Clin Cancer Res; 19(1); 194–204. ©2012 AACR.*



## Introduction

Lung cancer is the leading cause of cancer-related death worldwide (1). Even in early-stage patients treated by surgery, the risk of recurrence is high (2), and major efforts have been made to identify molecular markers that predict prognosis and response to additional therapy (3). Microarray-based gene expression profiling has successfully been used in clinical cancer research to subclassify cancer entities, to predict prognosis or response to therapy, and to identify underlying mechanisms of tumor development (4). In breast and colorectal cancer, prognostic gene expression signatures have been validated in independent patient cohorts and are now tested in prospective randomized clinical trials (5, 6).

Several prognostic gene expression signatures have been published in non–small cell lung cancer (NSCLC; refs. 7–17). However, a recent review critically evaluated the suggested signatures and concluded that in general they do not provide additional prognostic information compared with traditional clinical parameters (18). Some prognostic multigene signatures have been confirmed in independent datasets (14–16), but the impact of individual genes has rarely

**Authors' Affiliations:** Departments of [1]Immunology, Genetics and Pathology, [2]Surgical Sciences, and [3]Radiology, Oncology and Radiation Sciences, Section of Oncology, Uppsala University; [4]Regional Cancer Center Uppsala Örebro, Uppsala University Hospital, Uppsala; [5]Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm; Department of [6]Cardiothoracic Surgery and [7]Laboratory Medicine, Örebro University Hospital; [8]School of Health and Medical Sciences, Örebro University, Örebro, Sweden; [9]Department of Statistics, TU Dortmund University; [10]Leibniz Research Centre for Working Environment and Human Factors (IfADo), Dortmund, Germany; and [11]King's College London, Medical School, Division of Cancer Studies, London, United Kingdom

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

J. Botling and K. Edlund contributed equally to this work.

J. Rahnenführer, J.G. Hengstler, and P. Micke shared senior authorship.

**Corresponding Author:** Patrick Micke, Department of Immunology, Genetics and Pathology, Uppsala University, 751 85 Uppsala, Sweden. Phone: 46-18-6112615; Fax: 46-18-553354; E-mail: patrick.micke@igp.uu.se

**doi:** 10.1158/1078-0432.CCR-12-1139

## Translational Relevance

Gene signatures that predict survival in patients with NSCLC have been developed to identify patients with high risk of recurrence after radical resection and to personalize additional treatment options. However, the clinical value of multigene signatures is controversial, as they rarely outperform prognostication using conventional parameters. Also, the adaptation of multiplex mRNA-based assays to formalin-fixed paraffin-embedded (FFPE) tissue samples for clinical diagnostics remains a challenge. Instead, this study focused on the identification of single genes as prognostic biomarkers. A discovery screening was conducted in a novel well-characterized NSCLC cohort and candidate genes were confirmed in a meta-analysis of publicly available datasets. The presented genes showed prognostic impact and can be investigated further as promising biomarkers as well as targets for functional studies and drug development. Indeed, the potential application in routine diagnostics for one of these genes, CADM1, was verified by immunohistochemistry in two large tissue microarray cohorts of archived FFPE tissue samples.

been assessed. Surprisingly, there is virtually no overlap between hitherto published gene signatures. An explanation might be that effects of single genes with broad confidence intervals are difficult to confirm using a sequential validation strategy, that is, when genes identified as significant in one study are tested for significance in separate subsequent studies of comparably small sample size (19).

Rather than to find new gene signatures, the aim of this study was to evaluate the expression levels of single genes for prognostic relevance. To this end, we generated gene expression array data from a large well-characterized single-institute cohort of patients with operated NSCLC with complete clinical baseline information and long-term survival follow-up. We then sought to validate candidate genes in independent NSCLC datasets by the use of a meta-analysis approach in which the statistical significance associated with single genes is first assessed in each study separately. The significance across all studies is then calculated, combining the statistical power of multiple limited patient cohorts. As a proof-of-concept, our final goal was to test whether the prognostic impact of mRNA transcript levels were translated into protein expression differences that could be assessed as biomarkers by immunohistochemistry in routine clinical diagnostics.

## Materials and Methods

### Patients and tissue samples

The source population consisted of patients with surgically treated primary NSCLC, reported to the Uppsala-Örebro Regional Lung Cancer Registry consecutively from 1995 through 2005, with available fresh-frozen tissue in the Uppsala Biobank at the Department of Pathology. The Regional Lung Cancer Registry is a clinical audit and research database that prospectively compiles information such as diagnostic procedure, histology, stage, performance status according to WHO, smoking history, and survival for all diagnosed patients with lung cancer in the Uppsala-Örebro Region. The study was conducted in accordance with the Swedish Biobank Legislation and Ethical Review Act (Uppsala regional ethical review board, reference #2006/325 and Linköping regional ethical review board, reference #2010/44-31).

All fresh-frozen tissue samples were collected using the same standardized protocol (20–23), and each frozen section was reviewed by pathologists (P. Micke and J. Botling) to confirm that the sample contained representative tumor tissue. Study inclusion was based on: (i) NSCLC histology of squamous cell carcinoma, adenocarcinoma, or large cell carcinoma (including variants and NSCLC not otherwise specified), (ii) tumor sample size more than 5 mm, (iii) fraction of tumor cells 50% or more, (iv) follow-up time of more than 5 years, and (v) RNA integrity value (RIN) more than 7.0. If needed, tissue blocks were manually trimmed to enrich for tumor cells. Patients with a history of other cancers, or who had received neoadjuvant chemotherapy, were excluded. In total, 196 tissue samples met the inclusion criteria. All investigators involved in the study, apart from the study statistician, were blinded to patient outcome throughout all laboratory analyses.

Gene copy number data from 100 of these patients and global gene expression data from 78 patients have been described previously (GEO accession number GSE28582; ref. 24). For the present study, the histology of eight cases was reannotated. In addition, gene expression data from the study cohort ($n = 196$) were used to verify the prognostic impact of immunoglobulin κ C (25).

### RNA extraction and microarray analysis

Five to 10 sections (10 μm) were cut from each frozen tissue block and collected into a tube with Buffer RLT (Qiagen). Total RNA was extracted using the RNeasy Mini Kit following instructions from the manufacturer (Qiagen). RNA concentrations were measured with a ND-1000 spectrophotometer (NanoDrop Technologies), and the quality (RNA Integrity Number, RIN) was assessed using the Agilent 2100 Bioanalyzer system (Agilent Biotechnologies). For each sample, 2 μg of total RNA from each sample was used to prepare biotinylated fragmented cRNA for analysis on Affymetrix Human Genome U133 plus 2.0 arrays (54675 probe sets, Affymetrix Inc.). Sample preparation, processing, and hybridization were conducted according to the GeneChip Expression Analysis Technical Manual (Affymetrix Inc., Rev. 5). The arrays were washed and stained using a Fluidics Station 450 and finally scanned using a GeneChip Scanner 3000 7G. The subsequent analysis was carried out using the freely available statistical computing language R version 2.12.1 (http://www.r-project.org), including the R package meta for meta-analyses. The raw data (obtained as CEL files) were normalized using the robust multi-array average (RMA) method (26). The complete microarray

dataset has been deposited in the Gene Expression Omnibus database (GSE37745).

### Gene expression data analysis

Survival rates were calculated according to the Kaplan–Meier method. Overall survival (OS) was computed from the date of diagnosis to the date of death. Recurrence-free survival was computed from the date of diagnosis to the date of the last follow-up. Survival functions were compared with the log-rank test. Multivariate Cox survival analyses were conducted with inclusion of the most important clinical parameters. Categorization was conducted as follows: age: <70 versus ≥70 years; patient performance status 0 versus I–III, tumor stage I versus II–IV. In addition to clinical data collected by the regional lung cancer registry, a review of patient records provided information about recurrence-free survival for a subset of 96 patients, whereas data on adjuvant treatment could be reliably established for 100 patients.

For the meta-analysis, 5 publicly available gene expression datasets that used the Affymetrix gene chip U133A or U133 plus 2.0 arrays were included [Shedden and colleagues (27), Zhu and colleagues, GSE14814 (28), Raponi and colleagues, GSE4573 (29), Bild and colleagues, GSE3141 (30), Hou and colleagues, GSE19188 (31)] altogether comprising 860 patients with NSCLC (574 adenocarcinomas, 258 squamous cell carcinomas, and 28 large cell carcinomas or NSCLC not otherwise specified), with 22,277 probe sets overlapping between the arrays. The meta-analyses were conducted using the R package "meta" (http://CRAN.R-project.org/package = meta) with fixed effect models and random effects models based on parameter estimates of log HRs in Cox models and their SEs. For combining single estimates into one pooled estimate, inverse variance weighting was used. Significance of the overall effect was measured with the *P* value of the fixed effect models. Results were visualized with forest plots, also called confidence interval plots, in which parameter estimates of all single studies and the pooled estimates along with their confidence intervals are plotted on top of each other. Adjustment for multiple testing was conducted with the method of Benjamini and Hochberg (FDR; false discovery rate; ref. 32). All *P* values were 2-sided. Where no correction for multiple testing is indicated, the *P* values were considered as descriptive measures. The CADM1 metagene was constructed as previously described (33). All analyses were conducted using R version 2.12.1.

### Tissue microarray construction

The protein expression level of cell adhesion molecule 1 (CADM1) was analyzed in two independent cohorts using tissue microarrays (TMA) constructed from formalin-fixed paraffin-embedded tumor tissue. Hematoxylin and eosin–stained sections from all tissue blocks were reviewed by a pathologist to confirm the reported histologic subtype and to define representative tumor areas to be included in the array. The Uppsala cohort included 355 patients with NSCLC operated at the Uppsala University Hospital between 1995 and 2005 (34). Information on clinical patient parameters (survival time, histology, tumor stage, performance status, smoking history) was retrieved from the Uppsala-Örebro Lung Cancer Registry and is presented in Supplementary Table S1. The Örebro cohort included tumor tissue from 262 patients with NSCLC operated at the Örebro University Hospital (Supplementary Table S2; ref. 35).

### Immunohistochemistry

Four-micrometer sections were cut from the TMA blocks, mounted on adhesive slides, and baked in 60°C for 45 minutes, followed by deparaffinization in xylene, hydration in graded alcohols, and blocking for endogenous peroxidase in 0.3% hydrogen peroxide. A pressure boiler (decloaking chamber, Biocare Medical) was used for antigen retrieval, boiling the slides for 4 minutes at 125°C in Target Retrieval Solution (Dako). Automated immunohistochemistry was conducted using an Autostainer XL ST5010 (Leica Microsystems GmbH). The primary antibody CADM1 (Atlas antibodies, Sigma-Aldrich; S4945) was diluted 1:10,000 in UltraAb Diluent (Thermo Fisher Scientific) and incubated for 30 minutes at room temperature. The slides were then incubated with a secondary reagent anti-rabbit/mouse HRP-conjugated UltraVision (Thermo Fischer Scientific) for 30 minutes at room temperature. Following washing steps, using diaminobenzidine as a chromogen, the slides were developed for 10 minutes and counterstained with Mayer's hematoxylin for 5 minutes (Sigma-Aldrich). The slides were mounted with Pertex (Histolab AB) and scanned using the Aperio ScanScope XT (Aperio) for generation of high-resolution digital images used in the evaluation of immunostaining.

The staining intensity was manually annotated using a 4-grade scale: negative (0), weak (1), moderate (2), and strong (3). The fraction of stained tumor cells was scored as follows: no positive cells (0), 1%–5% (1), 6%–25% (2), 26%–75% (3), and 76%–100% (4). For all stainings, one score was set for the duplicate (Uppsala) or triplicate (Örebro) cores on the arrays representing the same tumor sample. The ordinal scores for intensity and for fraction of stained tumor cells were multiplied, obtaining values in the range 0 to 12. This score was further dichotomized (low: 0–4, high: 5–12). Kaplan–Meier plots were used to visualize difference in survival between these 2 groups. Statistical significance of differences was obtained with log-rank tests. Furthermore, univariate and multivariate Cox models were fitted.

## Results

### Baseline characteristics of the Uppsala cohort

Gene expression microarray analysis was conducted on RNA prepared from 196 fresh-frozen NSCLC samples that matched the predefined histologic and RNA quality criteria. Clinical data and overall survival were retrieved from the regional cancer registry, with follow-up times in the range of 5 to 15 years (Supplementary Fig. S1A, Table 1). Kaplan–Meier plots on survival in patient groups stratified by gender, tumor histology, tumor stage and patient

**Table 1.** Demographic and clinical characteristics of NSCLC patients included in the gene expression analysis

| | N (%) |
|---|---|
| All cases | 196 (100.0) |
| Sex | |
| Male | 107 (54.6) |
| Female | 89 (45.4) |
| Age at diagnosis | |
| ≤70 | 151 (77.0) |
| >70 | 45 (23.0) |
| Median (range) | 65 (39–84) |
| Smoking History | |
| Current smoker | 96 (49.0) |
| Ex smoker | 85 (43.4) |
| Never smoker | 15 (7.7) |
| Stage at diagnosis | |
| IA | 40 (20.4) |
| IB | 90 (45.9) |
| IIA | 6 (3.1) |
| IIB | 29 (14.8) |
| IIIA | 21 (10.7) |
| IIIB | 6 (3.1) |
| IV | 4 (2.0) |
| Histology | |
| Squamous cell carcinoma | 66 (33.7) |
| Adenocarcinoma | 106 (54.1) |
| Large cell carcinoma/NOS[a] | 24 (12.2) |
| WHO performance status | |
| 0 | 105 (53.6) |
| 1 | 75 (38.3) |
| 2 | 12 (6.1) |
| 3 | 4 (2.0) |
| Mean follow-up (months) | 59.9 |

[a]NOS = NSCLC not otherwise specified.

performance status are shown in Supplementary Fig. S1B–E. As expected, performance status and tumor stage represented the clinical factors that exhibited the strongest association with overall survival.

Unsupervised hierarchical cluster analysis of the gene expression array data revealed that tumor histology is the prominent denominator for the global mRNA transcript profile of NSCLC. Two main groups contained most of the adenocarcinomas and squamous cell carcinomas, respectively, based on clustering of all probe sets as well as on the 1,000 probe sets with the lowest signal to noise ratio (Fig. 1). The large cell carcinoma/NSCLC not otherwise specified (NOS) cases were scattered within the 2 main clusters.

### Gene expression: univariate survival analysis in the Uppsala dataset

A Cox regression model was applied to identify genes with prognostic relevance. When all probe sets ($n = 54{,}675$) in the total cohort of 196 cases were analyzed, no gene showed a significant prognostic impact after adjustment for multiple testing with a false discovery rate (FDR) of 5%. This was also true for the squamous cell cancer subgroup ($n = 66$). In adenocarcinoma ($n = 106$), only two annotated genes (SSR4 and FAM46C) were identified to be associated with survival with strictly adjusted $P$ values. Neither SSR4 nor FAM46C have yet been described as prognostic markers in NSCLC. FAM46C has been suggested to be involved in the regulation of protein translation, but its exact function is unknown (36). Signal sequence receptor delta (SSR4) encodes one subunit of the translocon-associated protein complex involved in the transport of proteins across the endoplasmic reticulum membrane (37).

### Validation of candidate genes in a meta-analysis of NSCLC datasets

The workflow of the analysis is illustrated in Fig. 2. As rigorous adjustment for multiple testing may exclude relevant genes, we tested the 450 probe sets with an unadjusted significance level of $P < 0.01$ (Uppsala cohort, all NSCLC, $n = 196$, Supplementary Table S3A) in a meta-analysis of 5 available gene expression datasets generated on the Affymetrix platform, including in total 860 patients (27–31). Of the 450 analyzed probe sets, 62 were significantly associated with survival in the meta-analysis (unadjusted $P < 0.01$) and 17 thereof remained significant after adjustment for multiple testing (FDR<1%), that is, it is likely that all 17 probes sets, representing 14 genes (Table 2), have a true prognostic impact.

The same strategy was applied on adenocarcinomas and squamous cell carcinomas separately. In the Uppsala adenocarcinoma dataset, 658 probe sets showed an unadjusted significance level of $P < 0.01$ (Supplementary Table S3B). Of these, 56 probe sets were also significant in the meta-analysis (unadjusted $P < 0.01$). Finally, 6 probe sets, corresponding to 4 genes, were confirmed with a FDR<1% (Table 3). In squamous cell carcinoma, only 122 probe sets were associated with survival in the Uppsala cohort (unadjusted $P < 0.01$), and the meta-analysis confirmed 2 probe sets (unadjusted $P < 0.01$), none of which could be confirmed with a FDR<1% (Supplementary Table S3C).

### Performance of candidate genes in the Uppsala NSCLC cohort

To illustrate the prognostic impact of each probe set in the Uppsala cohort, we conducted a Kaplan–Meier analysis with samples dichotomized into 2 groups with expression levels less than or equal to median and levels more than median for each respective probe set. Supplementary Figure S2A shows the results for the 17 significant probe sets (unadjusted $P < 0.01$ in Uppsala cohort and meta-analysis, FDR<1%) in the complete NSCLC cohort. The 6 significant probe sets in the adenocarcinoma subgroup are displayed in Supplementary Fig. S2B. To compare the impact of the identified genes in the complete NSCLC cohort with the

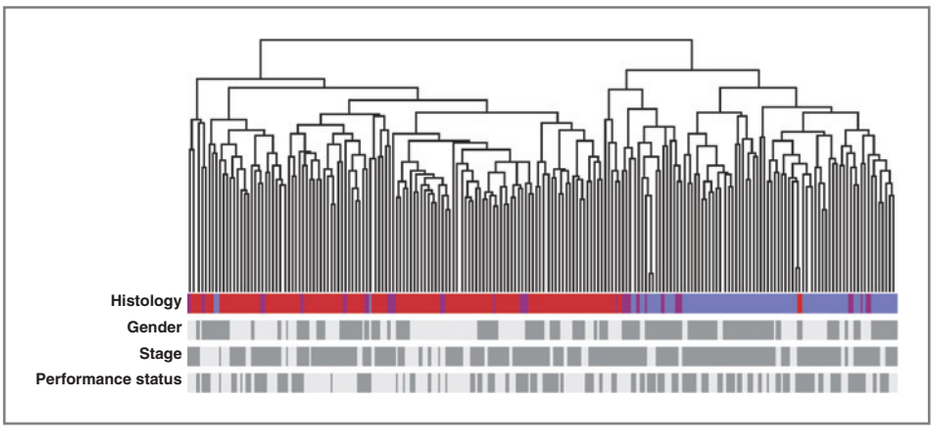


Figure 1. Unsupervised hierarchical cluster analysis. All NSCLC cases ($n = 196$) were clustered on the basis of the similarities of expression levels of the 1,000 genes with the lowest signal to noise ratio (mean divided by SD). The squares indicate histologic subtype (red = adenocarcinoma, blue = squamous cell carcinoma, purple = large cell carcinoma), gender (white = female, gray = male), stage (white = stage I, gray = stage II–IV), and performance status (white = PS 0, gray = PS I–IV) for each NSCLC sample.

most powerful clinical parameters (stage, age, performance status), we conducted univariate and multivariate Cox regression analyses of the candidate genes (Supplementary Table S4). All identified genes exhibited an independent prognostic association with survival.

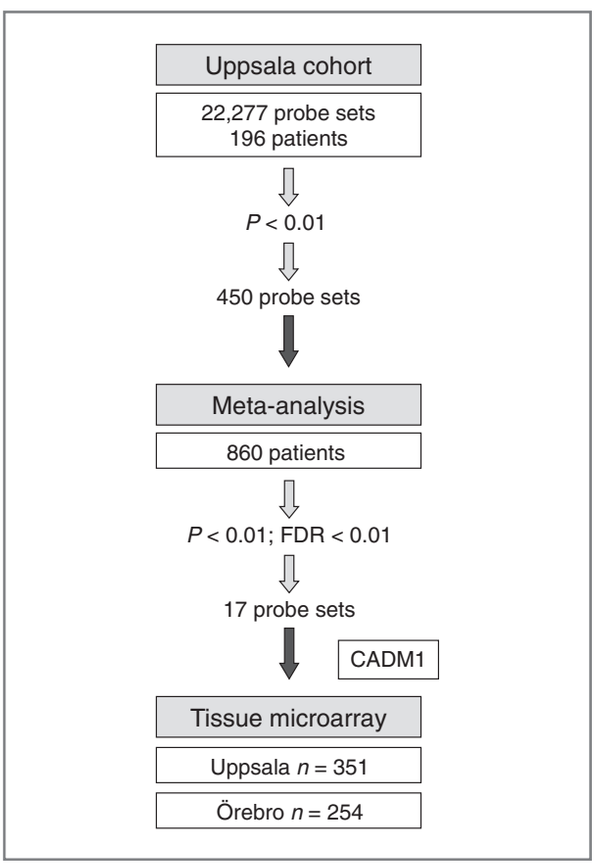


Figure 2. Flowchart of the applied analysis strategy. The prognostic impact of gene expression levels was analyzed in the Uppsala cohort. Probe sets that showed significance levels $P < 0.01$ were further tested in a meta-analysis. Sixty-two probe sets revealed significance levels $P < 0.01$, thereof 17 with a FDR < 1%. Of these genes, CADM1 was further analyzed by immunohistochemistry on 2 independent tissue microarrays.

## Higher CADM1 gene expression is associated with longer survival

As three CADM1 probe sets were identified as strong prognostic markers in the Uppsala cohort and in the meta-analysis, we constructed a CADM1 metagene of these probe sets (209030_s_at, 209031_at, 209032_s_at) and repeated the meta-analysis, now including the Uppsala cohort (Fig. 3A–C). Meta-analysis data indicated a significant prognostic impact of the CADM1 metagene in NSCLC, as well as in the adenocarcinoma and squamous cell carcinoma subgroups analyzed separately [all NSCLC: HR = 0.78; confidence interval (CI), 0.71–0.85; $P < 0.0001$; adenocarcinoma: HR = 0.74; CI, 0.66–0.83; $P < 0.0001$; squamous cell carcinoma: HR = 0.82; CI, 0.68–0.98; $P = 0.028$]. In the Uppsala cohort, the CADM1 metagene showed 2- and 5-year survival rates of 58% and 32% for patients with CADM1 expression below the median and 76% and 52% for patients with CADM1 expression above the median, respectively.

For a subset of patients, we were able to retrieve data with regard to recurrence-free survival and adjuvant chemotherapy. Longer recurrence-free survival showed a significant association with high CADM1 gene expression (Supplementary Fig. S3, $n = 96$; HR = 0.52; CI, 0.29–0.94; $P = 0.030$). When untreated patients were analyzed separately, high CADM1 gene expression was clearly associated with longer overall survival ($n = 71$; HR = 0.46; CI, 0.26–0.81; $P = 0.007$). For patients that had received adjuvant treatment, the prognostic impact was not significant ($n = 29$; HR = 0.67; CI, 0.29–1.54; $P = 0.34$) in patients with NSCLC with available clinical information. As the combined CADM1 probe sets displayed prognostic information independent of known prognostic markers, with a risk reduction of 43% in the Uppsala cohort (Table 4), we selected this gene for further immunohistochemical analysis.

## Immunohistochemical evaluation of CADM1 protein expression as a prognostic marker

A suitable antibody was identified using the Human Protein Atlas. Comprehensive protein profiling data, including underlying images showing the expression pattern of CADM1, are available through the Human Protein Atlas (www.proteinatlas.org; refs. 38, 39). CADM1 protein

**Table 2.** Probe sets and corresponding genes that were significantly associated (FDR<1%) with survival in the Uppsala cohort and meta-analysis including NSCLC of all histologies

| Affymetrix ID | Gene symbol | Gene name and function |
|---|---|---|
| 201037_at | *PFKP* | Phosphofructokinase, platelet |
| 202524_s_at | *SPOCK2* | Sparc/osteonectin, cwcv and kazal-like domains proteoglycan (testican) 2 |
| 202616_s_at | *MECP2* | Methyl CpG binding protein 2 |
| 204385_at | *KYNU* | Kynureninase (L-kynurenine hydrolase) |
| 210663_s_at | | |
| 205839_s_at | *BZRAP1* | Benzodiazapine receptor (peripheral) associated protein 1 |
| 206571_s_at | *MAP4K4* | Mitogen-activated protein kinase kinase kinase kinase 4 |
| 209030_s_at | *CADM1* | Cell adhesion molecule 1 |
| 209031_at | | |
| 209032_s_at | | |
| 211594_s_at | *MRPL9* | Mitochondrial ribosomal protein L9 |
| 217967_s_at | *FAM129A* | Family with sequence similarity 129, member A |
| 218092_s_at | *AGFG1* | ArfGAP with FG repeats 1 |
| 218451_at | *CDCP1* | CUB domain containing protein 1 |
| 218498_s_at | *ERO1L* | ERO1-like (*S. cerevisiae*) |
| 220658_s_at | *ARNTL2* | Aryl hydrocarbon receptor nuclear translocator-like 2 |
| 221497_x_at | *EGLN1* | Egl nine homolog 1 (*C. elegans*) |

expression was evaluated using immunohistochemistry on a tissue microarray (TMA) including 355 samples of operated NSCLC. This TMA includes formalin-fixed paraffin-embedded tissue from 189 of the 196 cases that were used for gene expression analysis, plus an additional 166 samples. In the end, 351 cases showed evaluable staining. High levels of membranous and/or cytoplasmic staining were observed in 90 samples, whereas low staining levels were shown in 261 cases (Fig. 4A and B).

Lower protein expression levels of CADM1 in the TMA cohort were significantly associated with shorter survival time in the univariate analysis (HR = 0.70; CI, 0.52–0.94; $P = 0.019$; Fig. 4A) as well as in the multivariate analysis including established clinical factors (HR = 0.69; CI, 0.51–0.93; $P = 0.016$; Supplementary Table S5). The 2-year survival rate was 81% for the high and 68% for the low expression group. The 5-year survival rates were 57% and 42% for patients with high and low CADM1 protein expression, respectively. Recurrence-free survival in patients with available clinical follow-up revealed a clear trend toward longer survival with CADM1 high staining ($n = 150$; HR = 0.60; CI, 0.34–1.06; $P = 0.076$).

In concordance with the gene expression analysis, the impact of CADM1 was more pronounced when the analysis was restricted to the adenocarcinoma subtype ($n = 195$, HR = 0.52; CI, 0.34–0.78; $P = 0.002$; Fig. 4B) with 2-year and 5-year survival rates 86% and 64% for high protein expression. For low protein expression, the 2- and 5-year survival rates were 70% and 40%, respectively. The association was not significant in the squamous cell carcinoma subtype ($n = 120$, $P = 0.80$). CADM1 was prognostic also when only patients with stage I adenocarcinoma ($n = 133$) were analyzed (Supplementary Fig. S4, HR = 0.53; CI, 0.33–0.88; $P = 0.013$). The 2-year and 5-year survival rates for patients with stage I adenocarcinoma with high CADM1 protein expression were 95% and 72% compared with 74% and 46% in the CADM1 low expression group, respectively. In patients with stage II adenocarcinoma the log-rank test did not reach significance, most probably because of the small sample size of only 30 cases ($P = 0.14$).

**Table 3.** Probe sets of genes that were significantly associated (FDR<1%) with survival in the Uppsala cohort and meta-analysis including only adenocarcinomas.

| Affymetrix ID | Gene symbol | Gene name |
|---|---|---|
| 200776_s_at | *BZW1* | Basic leucine zipper and W2 domain containing protein 1 |
| 201546_at | *TRIP12* | Thyroid hormone receptor interactor 12 |
| 209030_s_at | *CADM1* | Cell adhesion molecule 1 |
| 209031_at | | |
| 209032_s_at | | |
| 40148_at | *APBB2* | Amyloid beta (A4) precursor protein-binding, family B, member 2 |

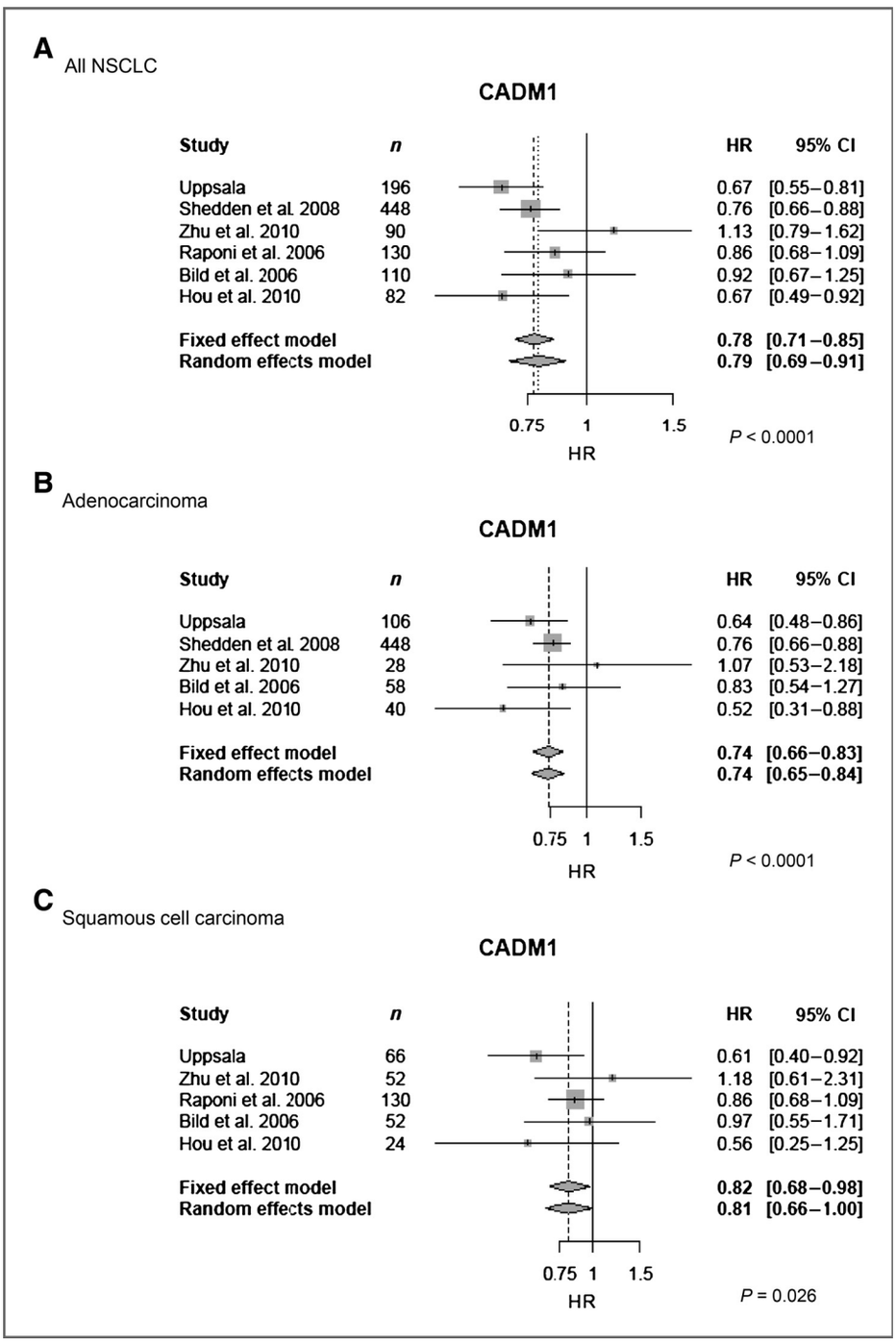


**Figure 3.** Meta-analysis of the prognostic impact of CADM1 gene expression. The three CADM1 probe sets were combined and included in a meta-analysis combining the Uppsala cohort and five publicly available NSCLC datasets. Results were presented as forest plots. The estimate from each study is represented by a gray box, in which the size of the box is proportional to the weight of the respective study in the meta-analysis. For each study, a horizontal line indicates the corresponding 95% CI. The plot is on a logarithmic scale so that the CIs are symmetric. A solid vertical line marks the HR 1 (no effect). The result of the meta-analysis (fixed effect and random effects model) is represented by a diamond, the center of the diamond indicates the pooled estimate (also marked by vertical lines), the dimension the corresponding 95% CI.

To further validate CADM1 staining as a diagnostic tool, we stained and evaluated an independent NSCLC cohort including 262 patients. The staining was evaluable in 254 cases. Patients with high CADM1 protein expression survived longer than patients with low protein expression (Fig. 5A; median survival 178 vs. 34 month, 2-year survival rate: 79% vs. 60%, 5-year survival rate: 61% vs. 43%, HR = 0.55; CI, 0.34–0.87; $P = 0.011$). Again the effect was more pronounced in patients with adenocarcinoma (Fig. 5B; median survival: not reached vs. 41 month; 2-year survival rate: 93% vs. 60%: 5-year survival rate: 77% vs. 45%, HR = 0.30; CI, 0.13–0.65; $P = 0.002$). Also in patients with stage I adenocarcinoma, the impact of CADM1 was significant ($n = 64$; HR = 0.22; CI, 0.05–0.94; $P = 0.042$). In conclusion, we were able to translate our findings from the array analysis of mRNA levels to clinically applicable immunohistochemical protein levels.

**Table 4.** Univariate and multivariate Cox regression model of 196 NSCLC cases including the CADM1 metagene (≤median vs. >median) and the most important prognostic parameters (dichotomized stage I vs. II–IV, age ≤70 vs. >70 years, performance status 0 vs. I–III)

| | Univariate | | Multivariate | |
|---|---|---|---|---|
| | HR (95% CI) | *P* value | HR (95% CI) | *P* value |
| CADM1 | 0.57 (0.41–0.79) | <0.001 | 0.66 (0.47–0.92) | 0.016 |
| Stage | 1.44 (1.03–2.02) | 0.033 | 1.42 (1.00–2.00) | 0.049 |
| Performance status | 1.91 (1.35–2.69) | <0.001 | 1.68 (1.20–2.34) | 0.002 |
| Age | 1.75 (1.21–2.54) | 0.003 | 1.65 (1.13–2.41) | 0.009 |

## Discussion

We introduced a novel NSCLC gene expression dataset with complete annotation of clinical parameters as well as long-term follow-up. The presented statistical approach was based on stringent criteria to evaluate the prognostic impact of single genes, and links our novel dataset to multiple independent patient cohorts. This combined screening and meta-analysis strategy identified with high confidence 17 probe sets (14 genes) independent of histology, and 6 probe sets (4 genes) in the subgroup of adenocarcinoma, that were associated with survival. As a proof-of-concept, we selected CADM1 for further evaluation. The prognostic impact of CADM1 was verified on the protein level in two independent cohorts, comprising altogether 605 evaluable NSCLC cases.

CADM1 belongs to the immunoglobulin superfamily (IGSF) of adhesion molecules and is located on chromosome 11q23. It was isolated primarily as IGSF4 and has been shown to function as a tumor suppressor in lung cancer (40); hence also named tumor suppressor in lung cancer 1 (TSLC1). Indeed, loss of heterozygosity (LOH) was observed in around 40% of lung cancers (41, 42) and loss of expression was also associated with hypermethylation of the corresponding promoter region in 44% of human lung cancers. In a previous study including 93 patients with NSCLC, loss of protein expression was associated with poor prognosis (43). In our study, we could confirm this result in two NSCLC cohorts using a novel polyclonal antibody. The staining intensity ranged from strong membranous staining to clear cut negative expression. These findings clearly support the clinical and biologic relevance of CADM1 as a tumor suppressor in NSCLC. In addition, we showed that the prognostic relevance is retained in patients with stage I disease. Thus CADM1 is a candidate marker for the stratification of patients for adjuvant chemotherapy. Because we used immunohistochemistry on formalin-fixed paraffin-embedded tissue, a direct implementation in diagnostics is possible.

Alongside CADM1, we present a list of additional candidate genes with a strong association with survival in NSCLC. Further studies are needed to characterize the functional and clinical importance of these genes. For instance, the

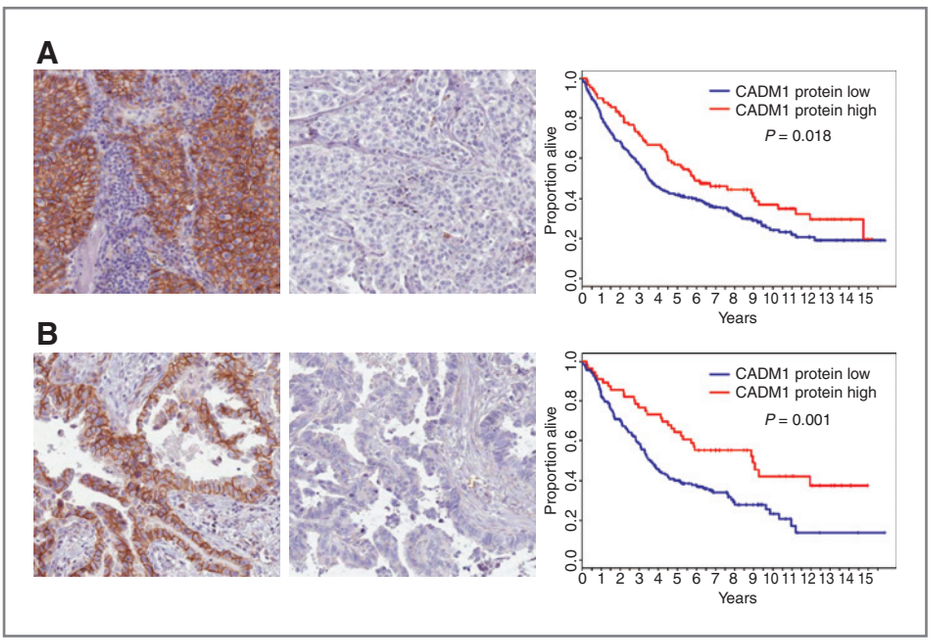


Figure 4. CADM1 protein expression in NSCLC. The tissue microarray, consisting of 351 evaluable NSCLC cases, was stained with a rabbit polyclonal antibody toward CADM1 and representative immunostainings are presented. The staining scores (0–12 as product of staining intensities and proportion of stained tumor cells) were dichotomized (0–4 vs. 5–12) and used for stratification in the Kaplan–Meier analysis. A, all NSCLC cases with available staining scores ($N = 351$). B, the adenocarcinoma subgroup ($n = 195$). The *P* values were obtained using a log-rank test.

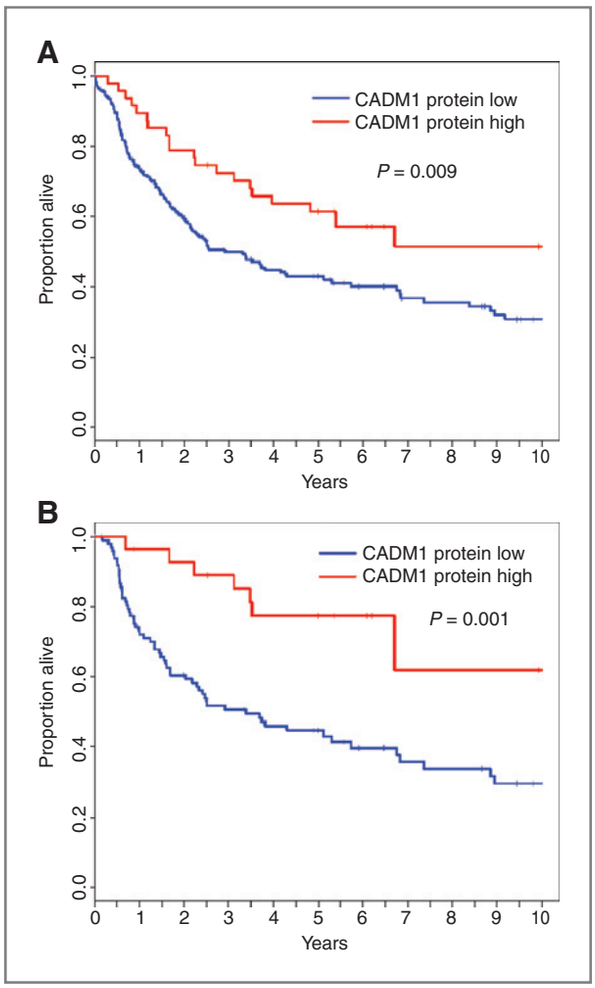


**Figure 5.** Kaplan–Meier analysis of CADM1 protein expression in the Örebro cohort. CADM1 protein expression was analyzed on a tissue microarray including tumor samples of 262 patients with NSCLC. The dichotomized staining intensity was used for stratification in the Kaplan–Meier analysis. A, all histological subtypes. B, adenocarcinoma only.

glycoprotein CUB domain containing protein 1 (CDCP1) represents a potential drugable target. CDCP1 is a cell surface protein, which has been linked to the EGFR, SRC, and AKT pathways and regulates PARP1-induced apoptosis. CDCP1 was described to be involved cancer migration and invasion, and inhibition effectively prevents tumor cell dissemination in animal models of prostate and lung cancer (44–46).

Previous studies have described prognostic gene expression signatures with varying discriminating prognostic power in independent datasets (7–17, 27–31). A clinical relevance of single genes included in these signatures has not yet been shown, and assays dependent on fresh-frozen tissue are difficult to introduce into routine diagnostics (18). Recently, a large-scale study applied a quantitative PCR-based assay for FFPE, including 11 cancer-related genes and 3 reference genes, to predict outcome in patients with operated nonsquamous NSCLC (47). The high-risk signature group showed 74%, the intermediate 57%, and the low risk group 45% five-year survival rates in a validation cohort of 1,006 Chinese patients. To evaluate this multigene signature in our cohort, we adjusted the expression values of the array data anticipating that they correlate with PCR data. As one gene (*WNT3*) was not represented on the Affymetrix array, the score is based on only 10 cancer target genes. With these limitations in mind, we analyzed the risk score in the Uppsala cohort and in the meta-analysis. Indeed, we could recapitulate previous findings, indicating that this multigene signature is robust (Supplementary Discussion). However, CADM1 gene expression and the immunohistochemical protein score displayed similar prognostic impact, indicating that comparable prognostic information can be obtained from single genes.

The strength of our study is that it is based on a large well-characterized NSCLC cohort, to our knowledge, one of the largest single-institute microarray datasets. Complete clinical annotation allowed analysis of clinically relevant subgroups, for example, stage I adenocarcinoma. Noteworthy is the inclusion of patient performance status, a parameter not annotated in previous datasets. Histopathologic review, fresh-frozen tissue handling, selection of representative tissue with high tumor cell content, and mRNA extraction were conducted using uniform and standardized protocols within the infrastructure of an established biobank and a diagnostic molecular pathology laboratory (20–23). Thus, we believe that artifacts due to poor tissue quality and methodologic inconsistencies have been minimized.

Applying sequential validation strategies, biomarkers identified in one study enter iteratively as candidate markers to be confirmed in other datasets. As a consequence, there is high confidence in the relevance of the final candidates. However, it has been shown that this approach yields many false negative results (19). Instead, we applied in this study a combined sequential and meta-analysis approach, in which candidate biomarkers with prognostic relevance in the primary cohort were further evaluated in a meta-analysis of several publicly available datasets. This procedure allows reliable validation in a large collection of samples by combining several independent smaller datasets with wide confidence intervals for single gene effects.

In conclusion, using a high quality NSCLC dataset together with an innovative meta-analysis approach, we identified novel prognostic genes with high reliability. On the basis of tissue microarray analysis of archived tissues, we showed the clinical relevance of CADM1 and a potential immunohistochemical application in molecular diagnostics.

## Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

## Authors' Contributions

**Conception and design:** J. Botling, L. Holmberg, M. Lambe, S. Ekman, J.G. Hengstler, P. Micke

**Development of methodology:** L. Holmberg, C. Karlsson, J. Rahnenführer

**Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.):** J. Botling, M. Lambe, S. Ekman, M. Bergqvist, F. Pontén, O. Fernandes, M. Karlsson, P. Micke

**Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis):** J. Botling, K. Edlund, M. Lohr, B. Hellwig,

L. Holmberg, A. Berglund, M. Bergqvist, A. König, C. Karlsson, J. Rahnenführer, J.G. Hengstler, P. Micke
**Writing, review, and/or revision of the manuscript:** J. Botling, K. Edlund, M. Lohr, L. Holmberg, M. Lambe, A. Berglund, S. Ekman, M. Bergqvist, A. König, O. Fernandes, M. Karlsson, C. Karlsson, J. Rahnenführer, P. Micke
**Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases):** K. Edlund, L. Holmberg, M. Lambe, M. Bergqvist, O. Fernandes, M. Karlsson, G. Helenius, P. Micke
**Study supervision:** J. Botling, P. Micke
**Histopathology review:** J. Botling, P. Micke

## Acknowledgments

The authors thank Anders Isaksson and Hanna Göransson-Kultima (Science for Life Laboratory, Department of Medical Sciences, Uppsala University, Uppsala, Sweden) for expert array service.

## Grant Support

This study was supported by the Swedish Cancer Society and Lions Cancer Foundation, Uppsala, Sweden and by the German Research Foundation (DFG, contract numbers RA 870/4-1 and RA 870/5-1). In addition, a part of this study was supported by an unrestricted research grant from Astra Zeneca. The Human Protein Atlas project was supported by grants from the Knut and Alice Wallenberg Foundation.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received April 11, 2012; revised September 11, 2012; accepted September 21, 2012; published OnlineFirst October 2, 2012.

## References

1. Jemal A, Siegel R, Xu J, Ward E. Cancer statistics, 2010. CA Cancer J Clin 2010;60:277–300.
2. Scott WJ, Howington J, Feigenberg S, Movsas B, Pisters K; American College of Chest Physicians. Treatment of non-small cell lung cancer stage I and stage II: ACCP evidence-based clinical practice guidelines (2nd edition). Chest 2007 132(Suppl):234S–42S.
3. Suzuki K, Kachala SS, Kadota K, Shen R, Mo Q, Beer DG, et al. Prognostic immune markers in non-small cell lung cancer. Clin Cancer Res 2011 17:5247–56.
4. Pusztai L. Chips to bedside: incorporation of microarray data into clinical practice. Clin Cancer Res 2006 12:7209–14.
5. Van Schaeybroeck S, Allen WL, Turkington RC, Johnston PG. Implementing prognostic and predictive biomarkers in CRC clinical trials. Nat Rev Clin Oncol 2011 8:222–32.
6. Reis-Filho JS, Pusztai L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. Lancet 2011 378: 1812–23.
7. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A 2001 98:13790–5.
8. Wigle DA, Jurisica I, Radulovich N, Pintilie M, Rossant J, Liu N, et al. Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. Cancer Res 2002 62:3005–8.
9. Tomida S, Koshikawa K, Yatabe Y, Harano T, Ogura N, Mitsudomi T, et al. Gene expression-based, individualized outcome prediction for surgically treated lung cancer patients. Oncogene 2004 23:5360–70.
10. Roepman P, Jassem J, Smit EF, Muley T, Niklinski J, van de Velde T, et al. An immune response enriched 72-gene prognostic profile for early-stage non-small-cell lung cancer. Clin Cancer Res 2009 15: 284–90.
11. Chen HY, Yu SL, Chen CH, Chang GC, Chen CY, Yuan A, et al. A five-gene signature and clinical outcome in non-small-cell lung cancer. N Engl J Med 2007 356:11–20.
12. Larsen JE, Pavey SJ, Passmore LH, Bowman RV, Hayward NK, Fong KM. Gene expression signature predicts recurrence in lung adenocarcinoma. Clin Cancer Res 2007 13:2946–54.
13. Lu Y, Lemon W, Liu PY, Yi Y, Morrison C, Yang P, et al. A gene expression signature predicts survival of patients with stage I non–small cell lung cancer. PLoS Med 2006;3:e467.
14. Guo NL, Wan YW, Tosun K, Lin H, Msiska Z, Flynn DC, et al. Confirmation of gene expression-based prediction of survival in non-small cell lung cancer. Clin Cancer Res 2008;14:8213–20.
15. Sun Z, Wigle DA, Yang P. Non-overlapping and non-cell-type-specific gene expression signatures predict lung cancer survival. J Clin Oncol 2008;26:877–83.
16. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med 2002;8:816–24.
17. Lee ES, Son DS, Kim SH, Lee J, Jo J, Han J, et al. Prediction of recurrence-free survival in postoperative non–small cell lung cancer patients by using an integrated model of clinical information and gene expression. Clin Cancer Res 2008;14:7397–404.
18. Subramanian J, Simon R. Gene expression-based prognostic signatures in lung cancer: ready for clinical use? J Natl Cancer Inst 2010;102:464–74.
19. Lohr M, Köllmann C, Freis E, Hellwig B, Hengstler JG, Ickstadt K, et al. Optimal strategies for sequential validation of significant features from high-dimensional genomic data. J Toxicol Environ Health A 2012; 75:447–60.
20. Micke P, Ohshima M, Tahmasebpoor S, Ren ZP, Östman A, Pontén F, et al. Biobanking of fresh frozen tissue: RNA is stable in nonfixed surgical specimens. Lab Invest 2006;86:202–11.
21. Botling J, Micke P. Biobanking of fresh frozen tissue from clinical surgical specimens: transport logistics, sample selection, and histologic characterization. Methods Mol Biol 2011;675:299–306.
22. Botling J, Micke P. Fresh frozen tissue: RNA extraction and quality control. Methods Mol Biol 2011;675:405–13.
23. Botling J, Edlund K, Segersten U, Tahmasebpoor S, Engström M, Sundström M, et al. Impact of thawing on RNA integrity and gene expression analysis in fresh frozen tissue. Diagn Mol Pathol 2009 18:44–52.
24. Micke P, Edlund K, Holmberg L, Kultima HG, Mansouri L, Ekman S, et al. Gene copy number aberrations are associated with survival in histologic subgroups of non–small cell lung cancer. J Thorac Oncol 2011 6:1833–40.
25. Schmidt M, Hellwig B, Hammad S, Othman A, Lohr M, Chen Z, et al. A comprehensive analysis of human gene expression profiles identifies stromal immunoglobulin kappa C as a compatible prognostic marker in human solid tumors. Clin Cancer Res 2012 Feb 20. [Epub ahead of print].
26. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003 19:185–93.
27. Shedden K, Taylor JM, Enkemann SA, Tsao MS, Yeatman TJ, Gerald WL, et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nat Med 2008 14:822–7.
28. Zhu CQ, Ding K, Strumpf D, Weir BA, Meyerson M, Pennell N, et al. Prognostic and predictive gene signature for adjuvant chemotherapy in resected non–small-cell lung cancer. J Clin Oncol 2010 28:4417–24.
29. Raponi M, Zhang Y, Yu J, Chen G, Lee G, Taylor JM, et al. Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. Cancer Res 2006 66:7466–72.
30. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 2006;439:353–7.
31. Hou J, Aerts J, den Hamer B, van Ijcken W, den Bakker M, Riegman P, et al. Gene expression-based classification of non-small cell lung carcinomas and survival prediction. PLoS ONE 2010;5:e10312.
32. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B 1995;57:289–300.
33. Schmidt M, Böhm D, von Törne C, Steiner E, Puhl A, Pilch H, et al. The humoral immune system has a key prognostic impact in node-negative breast cancer. Cancer Res 2008;68:5405–13.

34. Edlund K, Lindskog C, Saito A, Berglund A, Pontén F, Göransson-Kultima H, et al. CD99 is a novel prognostic stromal marker in non–small cell lung cancer. Int J Cancer 2012;131:2264–73.
35. Karlsson C, Helenius G, Fernandes O, Karlsson MG. Oestrogen receptor β in NSCLC—prevalence, proliferative influence, prognostic impact and smoking. Acta Pathol Microbiol Immunol Scand 2012;120:451–8.
36. Schoggins JW, Wilson SJ, Panis M, Murphy MY, Jones CT, Bieniasz P, et al. Diverse range of gene products are effectors of the type I interferon antiviral response. Nature 2011;472:481–485.
37. Brenner V, Nyakatura G, Rosenthal A, Platzer M. Genomic organization of two novel genes on human Xq28: compact head to head arrangement of IDH-gamma and TRAP-delta is conserved in rat and mouse. Genomics 1997;44:8–14.
38. Pontén F, Schwenk JM, Asplund A, Edqvist PH. The Human Protein Atlas as a proteomic resource for biomarker discovery. J Intern Med. 2011;270:428–46.
39. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, et al. Towards a knowledge-based Human Protein Atlas. Nat Biotechnol 2010;28:1248–50.
40. Gomyo H, Arai Y, Tanigami A, Murakami Y, Hattori M, Hosoda F, et al. A 2-Mb sequence-ready contig map and a novel immunoglobulin superfamily gene IGSF4 in the LOH region of chromosome 11q23.2. Genomics 1999;62:139–46.
41. Kuramochi M, Fukuhara H, Nobukuni T, Kanbe T, Maruyama T, Ghosh HP, et al. TSLC1 is a tumor-suppressor gene in human non–small-cell lung cancer. Nat Genet 2001;27:427–30.
42. Ito A, Okada M, Uchino K, Wakayama T, Koma Y, Iseki S, et al. Expression of the TSLC1 adhesion molecule in pulmonary epithelium and its down-regulation in pulmonary adenocarcinoma other than bronchioloalveolar carcinoma. Lab Invest 2003;83:1175–83.
43. Goto A, Niki T, Chi-Pin L, Matsubara D, Murakami Y, Funata N, et al. Loss of TSLC1 expression in lung adenocarcinoma: relationships with histological subtypes, sex and prognostic significance. Cancer Sci 2005;96:480–486.
44. Deryugina EI, Conn EM, Wortmann A, Partridge JJ, Kupriyanova TA, Ardi VC, et al. Functional role of cell surface CUB domain-containing protein 1 in tumor cell dissemination. Mol Cancer Res 2009;7:1197–211.
45. Casar B, He Y, Iconomou M, Hooper JD, Quigley JP, Deryugina EI. Blocking of CDCP1 cleavage *in vivo* prevents Akt-dependent survival and inhibits metastatic colonization through PARP1-mediated apoptosis of cancer cells. Oncogene 2011;31:3924–38.
46. Uekita T, Sakai R. Roles of CUB domain-containing protein 1 signaling in cancer invasion and metastasis. Cancer Sci 2011;102:1943–8.
47. Kratz JR, He J, Van Den Eeden SK, Zhu ZH, Gao W, Pham PT, et al. A practical molecular assay to predict survival in resected non-squamous, non-small-cell lung cancer: development and international validation studies. Lancet 2012;379:823–32.