Identification of transcriptional subgroups in EGFR-mutated and EGFR/KRAS-wild type lung adenocarcinoma reveals gene signatures associated with patient outcome

Maria Planck¹, Sofi Isaksson¹, Srinivas Veerla¹, Johan Staaf¹,²,*

¹Department of Oncology, Clinical Sciences, Lund University and Skåne University Hospital, Barngatan 2B, SE 22185 Lund, Sweden
²CREATE Health Strategic Center for Translational Cancer Research, Lund University, BMC C13, SE 22184, Lund, Sweden
*Corresponding author

Running title: Subgroups in EGFR-mutated and EGFRwt/KRASwt adenocarcinoma

Key words: lung cancer, EGFR, KRAS, prognosis, gene expression

Corresponding author:
Johan Staaf, e-mail: johan.staaf@med.lu.se, Address: Department of Oncology, Lund University, Barngatan 2B, SE 22185 Lund, Sweden, Fax: +4646147327

Competing interests
Authors declare that they have no competing interests.

Word count: 4441
Nbr figures: 4
Nbr tables: 1
Abstract

Purpose
In lung adenocarcinoma, EGFR and KRAS mutations dominate the mutational spectrum and have clear therapeutic implications. We sought to determine whether transcriptional subgroups of clinical relevance exist within EGFR-mutated, KRAS-mutated, or EGFR and KRAS-wild type (EGFRwt/KRASwt) adenocarcinomas.

Experimental Design
Gene expression profiles from 1186 adenocarcinomas, including 215 EGFR-mutated, 84 KRAS-mutated, and 219 EGFRwt/KRASwt tumors, were assembled and divided into four discovery (n=522) and four validation cohorts (n=664). Subgroups within the mutation groups were identified by unsupervised consensus clustering, SAM analysis and centroid classification across discovery cohorts. Genomic alterations in identified mutation subgroups were assessed by integration of genomic profiles for 158 cases with concurrent data. Multicohort expression subgroup predictors were built for each mutation group using the discovery cohorts, and validated in the four validation cohorts.

Results
Consensus clustering within the mutation groups identified reproducible transcriptional subgroups in EGFR-mutated and EGFRwt/KRASwt tumors, but not in KRAS-mutated tumors. Subgroups displayed differences in genomic alterations, clinicopathological characteristics, and overall survival. Multicohort gene signatures derived from the mutation subgroups added independent prognostic information in
their respective mutation group, for adenocarcinoma in general and stage I tumors specifically, irrespective of mutation status, when applied to the validation cohorts. Consistent with their worse clinical outcome, high-risk subgroups showed higher expression of proliferation related genes, higher frequency of copy number alterations/amplifications, and association with a poorly differentiated tumor phenotype.

Conclusions

We identified transcriptional subgroups in EGFR-mutated and EGFRwt/KRASwt adenocarcinomas with significant differences in clinicopathological characteristics and patient outcome, not limited to a mutation specific setting.

Translational relevance

EGFR and KRAS mutations dominate the mutational spectrum of lung adenocarcinoma. EGFR mutation is an established predictive marker for response to targeted therapy, although primary or acquired resistance impairs the result of treatment with EGFR inhibitors. Further characterization of mutation groups may identify new molecular subgroups, additional targets for synergistic treatment, and provide new insights into resistance mechanisms and molecular pathogenesis. Based on a multicohort discovery and validation strategy, we identified transcriptional subgroups in EGFR-mutated and EGFR/KRAS-wild type adenocarcinomas with significant differences in genomic alterations, clinicopathological characteristics, and prognosis. Moreover, these subgroup gene signatures also added independent prognostic information in adenocarcinomas in general and in stage I disease.
specifically, irrespective of mutation status. Further investigations on the predictive value of these gene signatures in the setting of targeted treatment may provide a future basis for refined diagnosis and treatment of lung adenocarcinoma.

Introduction

Lung cancer is the leading cause of cancer-related mortality worldwide (1). The disease is heterogeneous but may be broadly divided into small cell lung cancer and non-small cell lung cancer (NSCLC). NSCLC accounts for ~85% of all diagnosed cases, with adenocarcinoma as the most frequent histological type (2). The epidermal growth factor receptor gene, \textit{EGFR}, located at 7p11.2 and the V-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog gene, \textit{KRAS}, located at 12p12.1 represent the two most frequently mutated oncogenes in adenocarcinoma (3). \textit{EGFR} and \textit{KRAS} mutations are essentially mutually exclusive in these tumors and are associated with differences in e.g. patient gender and smoking history, suggesting that these molecular defects may be drivers of pathogenesis for specific subgroups (3) (and references therein). In line with this, \textit{EGFR} mutations have been associated with improved overall survival, whereas \textit{KRAS} mutations may predict shorter survival for patients with advanced lung adenocarcinoma (4). Furthermore, the occurrence of \textit{EGFR} mutations predicts an improved response to EGFR tyrosine kinase inhibitors and is therefore routinely assessed in clinical practice (5).

\textit{EGFR} and \textit{KRAS}-wild type (EGFRwt/KRASwt) adenocarcinomas represent a still unclear group, with different potential driver mutations as well as mutually exclusive genomic rearrangements of the \textit{ALK}, \textit{RET}, and \textit{ROS1} genes (3, 6, 7). Interestingly, a subgroup of EGFRwt/KRASwt adenocarcinomas appears to benefit
from EGFR inhibitors, although there are currently no predictive markers to identify these patients (8, 9). Consequently, increased knowledge about the molecular background of tumor subgroups defined by EGFR/KRAS mutation status may identify novel genomic signatures and additional targets for synergistic treatment, and also provide new insights into resistance mechanisms and molecular pathogenesis.

Numerous single biomarkers (reviewed in, e.g. (10, 11)) and microarray-based gene signatures (12-16) associated with clinical outcome in NSCLC/adenocarcinoma have been reported to date. In addition, different molecular subgroups in lung adenocarcinoma have also been reported (17, 18), of which the bronchioid, squamoid, and magnoid subtypes originally defined by Hayes et al. (17) have been reproduced in multiple cohorts (17, 19). Bronchioid tumors are generally of lower grade, have a higher proportion of EGFR mutations and higher expression of excretion, asthma and surfactant genes, occur predominantly in women and never-smokers, and have better overall survival (17, 19). In contrast, magnoid and squamoid tumors harbor more KRAS mutations, seem to be more closely related in gene expression, occur more often in men and smokers, and have poorer overall survival (17, 19). Despite these findings, the clinical usefulness of many of these markers/signatures remains debated (11, 20), and the prognostic performance in adenocarcinoma subgroups defined by EGFR and KRAS mutational status remains relatively unclear. Several studies have also illustrated the difficulties in separating EGFR-mutated, KRAS-mutated and EGFRwt/KRASwt tumors into distinct transcriptional entities (16, 21-23). Moreover, it has not been systematically investigated whether clinically relevant transcriptional subgroups exist within the mutation groups.

Herein, we sought to determine whether transcriptional subgroups of clinical relevance exist within EGFR-mutated, KRAS-mutated, and/or EGFRwt/KRASwt
adenocarcinomas. Through a multicohort discovery and validation strategy we identified reproducible gene expression subgroups within \textit{EGFR}-mutated tumors and \textit{EGFR}wt/\textit{KRAS}wt tumors associated with different genomic alterations, clinicopathological characteristics, and patient outcomes. In addition, these subgroup gene signatures also added independent prognostic information in adenocarcinomas in general and in stage I disease specifically, irrespective of mutation status.

**Materials and Methods**

*Tumor cohorts*

Gene expression profiles from 522 lung adenocarcinomas, including 215 \textit{EGFR}-mutated, 84 \textit{KRAS}-mutated and 219 \textit{EGFR}wt/\textit{KRAS}wt cases, were obtained from GSE31210 (16), E-MTAB-923 (21) and Chitale et al. (22) and were used as discovery cohorts. Samples from Chitale et al. (22) were divided into two cohorts based on the different Affymetrix platforms used in this study (U133A and U133 2plus) creating four final discovery cohorts (Table 1). Prognostic associations of the derived gene signatures were validated in 117 tumors with known mutation status from GSE13213 (15) (n=45 \textit{EGFR}-mutated, 15 \textit{KRAS}-mutated and 57 \textit{EGFR}wt/\textit{KRAS}wt), and 547 additional adenocarcinomas with unknown mutation status from Shedden et al. (14) (n=356), GSE3141 (24) (n=58), and the University of Texas Lung Specialized Program of Research Excellence cohort (13) (UT Lung SPORE, GSE42127, n=133 adenocarcinomas). Matched and analyzed genomic profiles from 158 tumors (53 \textit{EGFR}-mutated and 105 \textit{EGFR}wt/\textit{KRAS}wt) belonging to the E-MTAB-923 and Chitale et al. cohorts were extracted from Staaf et al. (25).
Response to adjuvant chemotherapy for patients with NSCLC was investigated in the complete UT Lung SPORE cohort (GSE42127, n=176) including patients treated mainly with carboplatin plus taxanes, and in the National Cancer Institute of Canada Clinical Trials Group JBR.10 clinical trial cohort (n=90, GSE14814) (12) including patients treated mainly with vinorelbine plus cisplatin (Table S1). For both of these cohorts, gene expression profiling was performed prior to therapy. Response to sorafenib, a drug that targets different tyrosine and Raf kinases, treatment of NSCLC patients with advanced chemorefractory metastatic disease was explored in the GSE33072 cohort (26) (Table S1). The biopsy samples in this study were taken from the lung, liver, lymph node, bone/soft tissue, and adrenal glands of patients enrolled in the Biomarker-integrated Approaches of Targeted Therapy for Lung Cancer Elimination (BATTLE) trial (27).

Explicit information on patient ethnicity or specific mutation type was not available for the majority of included patients, and was therefore omitted from analyses. Included studies were performed in both western and Asian countries. Patient and tumor characteristics are summarized in Table 1 and Table S1.

Gene expression analyses

Affymetrix cohorts were normalized using GC Robust Multi-array Averaging (GCRMA) (28), except for GSE3141 (24) and GSE33072 (26) for which normalized expression data was obtained from Gene Expression Omnibus (29). Normalized expression data was obtained from Gene Expression Omnibus for non-Affymetrix cohorts. Unsupervised subgroup discovery within mutation groups was performed using consensus clustering through the ConsensusClusterPlus R package (30) on the four discovery cohorts individually. Significance Analysis of Microarrays (SAM) (31)
from the siggenes R package (32) was used to identify differentially expressed probe sets between consensus clusters for the discovery cohorts. Probe sets with false discovery rate <5% were considered statistically significant. Nearest-centroid predictors and multicohort centroids were created for each cohort or mutation group as described (Supplementary Information). Proliferation differences between samples were assessed by an expression metagene based on the proliferation/chromosome instability (CIN70) signature (33) (referred to as the CIN70 metagene hereon), which includes numerous proliferation/cell cycle-related genes (Supplementary Information). Data processing steps are further described in Supplementary Information.

Survival analysis

Survival analyses were performed in R using the Survival package with overall survival as endpoint. Survival curves were compared using Kaplan-Meier estimates and the log-rank test. Follow-up time for overall survival was censored at 5 years for all cohorts.

Results

Unsupervised discovery of reproducible transcriptional subgroups in EGFR-mutated and EGFRwt/KRASwt tumors

Figure 1 shows a schematic of the steps taken to identify and validate transcriptional subgroups within the three mutation groups (see also Supplementary Information). For each cohort and mutation group, consensus clustering was used to identify two
sample clusters (Figure 2) from which a centroid classifier was constructed based on differentially expressed genes from a SAM analysis. Next, each mutation and cohort-specific classifier was used to classify tumors of similar mutation status in the remaining three discovery cohorts, and the overlap between the predicted groups and the original consensus clusters was compared for each classifier and for each cohort.

This cross-cohort approach identified two reproducible subgroups in \textit{EGFR}-mutated tumors, termed EGFR-1 and EGFR-2 herein, based on consistency between predicted and original consensus clusters in three out of four discovery cohorts (Figures 2A and S1A). Similarly, two subgroups, termed wt/wt-1 and wt/wt-2 hereon, were identified in EGFRwt/KRASwt tumors based on all four discovery cohorts (Figures 2B and S1B). By contrast, no robust subgroups (across at least three discovery cohorts) were identified in \textit{KRAS}-mutated tumors (data not shown). For EGFRwt/KRASwt tumors the wt/wt-1 and wt/wt-2 subgroups in GSE31210 corresponded strongly to the clusters identified by Okayama et al. (16); 95% of the tumors were similarly grouped (Fisher’s exact test \(p=2\times10^{-17}\)). Notably, for both EGFR-1/2 and wt/wt-1/2, large transcriptional differences between groups were observed in the SAM analysis, comprising thousands of probe sets (Figures S1A and B).

\textit{EGFR-mutated and EGFRwt/KRASwt transcriptional subgroups are associated with clinicopathological differences, molecular subtypes and patient outcome}

Across the discovery cohorts, \textit{EGFR}-mutated and EGFRwt/KRASwt transcriptional subgroups were associated with differences in a) adenocarcinoma molecular subtype (19) patterns, b) expression of a proliferation metagene (CIN70) (33), c) clinicopathological characteristics, and d) overall survival. In \textit{EGFR}-mutated tumors,
the EGFR-1 subgroup included 75-85% of all bronchioid but few magnoid or squamoid classified tumors (19), displayed lower expression of the CIN70 metagene, included older patients, and displayed better overall survival compared with EGFR-2 (Figures 3A and S1C). No associations with tumor stage (p<0.05 in GSE31210, Fisher’s exact test), patient gender (p>0.05 all cohorts, Fishers’s exact test) or smoking status (p<0.05 in GSE31210, Fisher’s exact test) were observed in more than one discovery cohort for EGFR-1 or EGFR-2, potentially due to the small group sizes for certain cohorts. In EGFRwt/KRASwt tumors the wt/wt-2 subgroup included 88-100% of the bronchioid tumors but few squamoid/magnoid tumors, displayed lower expression of the CIN70 metagene, included more never-smokers, and displayed better overall survival (with the exception of E-MTAB-923) compared with wt/wt-1 (Figures 3B and S1D). No associations with patient gender, age, or tumor stage were observed for wt/wt-1 or wt/wt-2 in the discovery cohorts, potentially due to the small group sizes for certain cohorts (p>0.05 all cohorts, Fisher’s exact test or Wilcoxon’s test).

EGFR-mutated and EGFRwt/KRASwt transcriptional subgroups display different genomic alterations

In 158 tumors with matched genomic and transcriptional profiles the high-risk subgroups, EGFR-2 and wt/wt-1, were both associated with overall more copy number alterations in their respective mutation group (measured as the fraction of the genome altered by copy number alterations, see Supplementary Information and (25)) compared with their corresponding low-risk subgroups (p=0.04, and p=0.0007, respectively, Wilcoxon’s test). EGFR-mutated EGFR-2 tumors displayed more copy number gain in regions on 7p (including \textit{EGFR}, ~85% of cases) and 3q, and more
frequent losses on 4q, 9p, 15q and 16q compared with EGFR-mutated EGFR-1 tumors (Figure S2A, >25% frequency difference). Moreover, 80% of the EGFR amplifications, 89% of all 7p amplifications, and all NKX2-1/TITF amplifications were found in the EGFR-2 subgroup compared with EGFR-1 in the analyzed EGFR-mutated tumors.

In EGFRwt/KRASwt tumors, the high-risk wt/wt-1 subgroup displayed more copy number gains in regions on 1q, 3q, 5p, 7q and 12p and more frequent losses on 4q, 5q, 15q and 22q compared with the wt/wt-2 subgroup (Figure S2B, >25% frequency difference). Overall, wt/wt-1 tumors also displayed more amplifications compared with wt/wt-2 (p=0.0003, Fisher’s exact test). Two recurrently amplified regions were found to differ significantly between the wt/wt-1 and wt/wt-2 subgroups: 8p12 (FGFR1) amplifications in wt/wt-1, and 12q15 (MDM2) amplifications in wt/wt-2 (p=0.02 and p=0.03, respectively, Fisher’s exact test).

Independent validation of EGFR-mutated and EGFRwt/KRASwt transcriptional subgroups

To validate the EGFR-mutated and EGFRwt/KRASwt transcriptional subgroups in independent cohorts we first created a single EGFR-mutated (EGFR-1/2) and a single EGFRwt/KRASwt (wt/wt-1/2) multicohort centroid classifier from the individual discovery cohort centroids for respective mutation group (Figure 1, Supplementary Information, and Table S2). Notably, 18-20% of the genes in the EGFR-1/2 and wt/wt-1/2 multicohort classifiers matched reported lists of potential therapeutic targets and modulators of chemotherapy drugs’ effects in lung cancer cells (34, 35) (Table S2), whereas 24-26% of the genes in the two signatures overlapped with the
bronchioid, magnoïd and squamoid subtype centroids reported by Wilkerson et al. (19).

The two multicohort centroid classifiers were next applied to their respective mutation group in the GSE13213 (15) cohort. Consistent patterns of molecular subtype distribution, CIN70 metagene expression, patient age, smoking status, and overall survival, were observed for the predicted subgroups in GSE13213 compared with the discovery cohorts, although not always reaching statistical significance due to small sample sizes (Figures 3C-D and S1E-F). Consistent with the largest EGFR-mutated discovery cohort (GSE31210), the low-risk EGFR-1 subgroup in GSE13213 contained nearly twice as many never-smokers as the EGFR-2 subgroup (Figure S1E-F). Classification of all 117 samples in GSE13213 showed that both multicohort signatures were associated with overall survival (log-rank p=2x10^{-5} for EGFR-1/2 and p=2x10^{-4} for wt/wt-1/2), and that there was a high consistency between EGFR-1/2 and wt/wt-1/2 classifications (p=1x10^{-21}, Fisher’s exact test). Both classifications also added independent prognostic information in multivariate analysis including tumor stage, patient age, mutation status, and smoking status as covariates, and overall survival as the endpoint (Figures 4A-B).

To further validate the general prognostic association of the two multicohort signatures we classified 547 additional independent adenocarcinomas with unknown mutation status from Shedden et al. (14), GSE3141 (24), and GSE42127 (13) (Table 1). Consistently, centroid classifications were associated with overall survival in Shedden et al. (log-rank p=4x10^{-5} for EGFR-1/2 and p=9x10^{-5} for wt/wt-1/2), GSE3141 (p=0.01 and p=0.006, respectively) and GSE42127 (p=0.008 and p=0.0006, respectively). In Shedden et al., the high-risk EGFR-2 and wt/wt-1 groups were both strongly enriched with poorly differentiated tumors, while the corresponding low-risk
EGFR-1 and wt/wt-2 groups included >90% of the well-differentiated tumors in this cohort (p=6x10^{-20} and p=4x10^{-19}, respectively, Fisher’s exact test). In multivariate analysis both multicohort classifications added independent prognostic information in GSE42127 with tumor stage, patient age, and adjuvant chemotherapy as covariates, and in Shedden et al. when including tumor stage, patient age, adjuvant chemotherapy, and smoking status as covariates (Figures 4A-B). Furthermore, both classifications also added independent prognostic information for patients with stage I disease in Shedden et al., GSE42127, and GSE13213 (Figures 4C-D).

Patients with bronchioid classified tumors have repeatedly been shown to have superior outcomes to patients with the magnoid or squamoid subtypes (17, 19). Considering the enrichment of bronchioid classified tumors in the low-risk groups, we investigated whether the multicohort classifiers added independent prognostic information compared with the molecular subtypes by including the molecular subtypes in the previous multivariate models for Shedden et al., GSE13213, and GSE42127. In GSE42127, neither the EGFR-1/2 (p=0.37), the wt/wt-1/2 (p=0.07), nor the molecular subtype classifications were significant in the multivariate analysis. In the Shedden et al. cohort both multicohort classifiers added independent prognostic information in multivariate analysis (n=221 EGFR-2: p=0.01, HR=2.2, 95% Confidence interval (CI) =1.2-4, and n=220, wt/wt-1: p=0.03, HR=1.9, 95% CI=1.1-3.3). Similarly, in GSE13213, both multicohort classifiers also added independent prognostic information after inclusion of molecular subtype in the multivariate model (n=117 EGFR-2: p=0.0002, HR=10, 95% CI=3-34, and n=117, wt/wt-1: p=0.045, HR=4.3, 95% CI=1.03-20). Similar multicohort centroid classification of squamous cell carcinoma tumors in GSE3141 (n=53) or GSE42127 (n=43) was not associated
with overall survival (log-rank p>0.05, data not shown), suggesting that the prognostic association of the multicohort signatures is adenocarcinoma specific.

*EGFR-mutated and EGFRwt/KRASwt gene signatures are associated with response to treatment in NSCLC*

To assess whether the EGFR-1/2 and wt/wt-1/2 signatures were associated with response to treatment we first analyzed gene expression profiles from tumor biopsies taken from NSCLC patients with advanced chemorefractory metastatic disease enrolled in the BATTLE trial (26, 27) (GSE33072, Table S1). We restricted the analysis to the 30 EGFRwt/KRASwt NSCLC patients treated with sorafenib due to otherwise small sample groups and poor sample annotations. For the sorafenib-treated cohort the high-risk EGFR-2 and wt/wt-1 groups included the majority of patients who did not meet the primary endpoint of 8-week disease control (85% and 83%, respectively). Moreover, the wt/wt-1 classification was borderline non-significant for worse progression-free survival compared with wt/wt-2 (log-rank p=0.07).

Secondly, we investigated whether the multicohort classifiers predicted survival benefits from adjuvant chemotherapy (ACT) in NSCLC based on the GSE14814 (JBR.10 clinical trial cohort, n=90) (12) and GSE42127 (UT Lung SPORE, n=176) (13) cohorts (Figure 1, Table S1). In GSE14814, ACT-treated NSCLC patients showed significantly better overall survival than those without treatment in the high-risk EGFR-2 (n=58, HR=0.39; 95% CI=0.16-0.97; p=0.04) and wt/wt-1 (n=56, HR=0.39; 95% CI=0.15-0.99; p=0.05) groups, whereas in the low-risk groups ACT had no significant survival benefits. Similar results were found for the GSE42127 cohort, with NSCLC patients in the high-risk groups benefitting from ACT (EGFR-2 n=98, HR=0.42; 95% CI=0.17-1; p=0.05 and trend-like for wt/wt-1
n=83, HR=0.46; 95% CI=0.18-1.2; p=0.11), while patients in the low-risk groups did not benefit from ACT.

**Discussion**

In this study we investigated whether reproducible transcriptional subgroups could be identified within *EGFR*-mutated, *KRAS*-mutated and/or *EGFRwt/KRASwt* lung adenocarcinomas based on a multicohort discovery and validation approach. Unsupervised subgroup discovery within individual mutation groups identified transcriptional subgroups in *EGFR*-mutated and *EGFRwt/KRASwt* tumors associated with differences in genomic alterations, clinicopathological characteristics, and patient outcome. The failure to identify reproducible subgroups within *KRAS*-mutated tumors could be due to the lower number of available tumors and/or a potential biological or etiological heterogeneity among these tumors (36). For *EGFRw/KRASwt* tumors our results extend recent findings by Okayama et al. (16), by demonstrating that apparently similar subgroups exist in other cohorts using a different analysis approach. Subgroup signatures included a spectrum of genes involved in lung carcinogenesis (*NKX2-1*, *HOPX*, *LOXL2*, several matrix metallopeptidases), DNA-repair (*BRCA1*, *RAD51*, *MSH6*, *MSH2*), cell proliferation and cell cycle control, as well as genes coding for secretory proteins and collagens (see Table S2 and Figure S3).

Previous studies on resected lung adenocarcinoma have suggested a large number of gene signatures, single biomarkers (including *ERCC1*, *RRM1*, and different cell cycle regulators), and molecular subtype signatures to be associated with survival (see e.g. (10-18)). With respect to these, our study adds a demonstration of how
unsupervised analysis of transcriptional patterns in straightforwardly defined, and clinically relevant adenocarcinoma mutation groups can stratify patients into better or worse prognosis in both a mutation-specific and a general setting. Importantly, cohorts included in both the discovery and validation phases were analyzed by different microarray platforms and performed in both western and Asian countries. In order to develop a more clinically practical molecular assay the multicohort gene signatures would need to be reduced to a smaller set of marker genes, e.g. by gene network analyses as recently described (13, 37). Such a reduced gene set may be measured by other techniques with potential application also to formalin-fixed paraffin embedded tissue.

The prognostic associations of the multicohort EGFR-1/2 and wt/wt-1/2 gene signatures were adenocarcinoma specific, but not limited to a mutation specific context, since both signatures added independent prognostic information irrespective of mutation status in four independent adenocarcinoma validation cohorts. The association with patient outcome in adenocarcinomas for the two signatures is presumably due to the presence of a strong proliferative component, with elevated cell proliferation and loss of cell cycle control associated with poor outcome (14, 38). The prognostic importance of the proliferative component in the signatures was supported by the observation that overlapping probe sets between the EGFR-1/2 and the wt/wt-1/2 multicohort signatures were strongly enriched for cell cycle-related genes. In addition, a centroid classifier based on these overlapping probe sets alone yielded nearly identical prognostic results as the original classifiers (data not shown).

The EGFR-1 and wt/wt-2 low-risk groups were notably enriched for bronchioid classified tumors, while the high-risk groups included the majority the of magnoind and squamoid tumors. Bronchioid classified tumors have been repeatedly
associated with EGFR alterations (17, 19), although ~30% or more of the EGFR-mutated tumors are classified as non-bronchioid in discovery cohorts from both previous studies (17, 19) and the current study. In the absence of bronchioid tumors we found no significant association between the magnoid and squamoid subtypes and EGFR/KRAS mutation status in our four discovery cohorts and GSE13213. Although a detailed analysis of the characteristics of the molecular subtypes with respect to EGFR/KRAS mutation-defined subgroups have not been reported, proliferation differences appear as a possible explanation for the differences in distribution of the bronchioid, magnoid and squamoid subtypes between our mutation subgroups. This seems likely since bronchioid tumors overall displayed significantly lower expression of the CIN70 metagene compared with magnoid and squamoid tumors in all discovery cohorts irrespective of whether tumors were stratified by mutation status or not (data not shown). In contrast, no significant difference was seen in expression of the CIN70 metagene between magnoid and squamoid tumors (data not shown). In line with this, increasing the number of evaluated clusters in the consensus clustering from two to three for the largest cohorts of EGFR-mutated tumors (GSE31210, n=127 and E-MTAB-923, n=49) or EGFRwt/KRASwt tumors (GSE31210, n=79 and Chitale U133A, n=65) did not resolve magnoid and squamoid tumors as separate clusters (data not shown). In addition, proliferation differences between mutation subgroups also appear as the likely explanation for the enrichment of never-smokers in the low-risk wt/wt-2 subgroup and the low-risk EGFR-1 group in GSE31210 and GSE13213 (39, 40).

In recent years, adjuvant chemotherapy has improved the overall survival for surgically treated NSCLC patients (41), and there are a few reports of gene signatures that predict response to adjuvant treatment (12, 13, 42). Gene expression-based
classification of NSCLC cases in the GSE14814 (12) and GSE42127 (13) cohorts by the EGFR-1/2 and wt/wt-1/2 signatures in our study suggested that adjuvant chemotherapy only benefits patients in the high-risk EGFR-2 and wt/wt-1 groups. Notably, these results are consistent with previous studies reporting predictive gene signatures based on the same cohorts (12, 13, 42). However, these results should be interpreted with care given the small cohort sizes and the increasing evidence that tumor histology needs to be considered in clinical decision making for treatment of NSCLC (43), as exemplified by the superior effect of pemetrexed in non-squamous NSCLC (44). For instance, in both the GSE14814 and the GSE42127 cohorts adjuvant chemotherapy appeared more beneficial in a 5-year follow-up perspective for patients with squamous cell carcinomas (p=0.06 and p=0.10, respectively, log-rank test) compared with patients with adenocarcinomas (p=0.21 and p=0.32, respectively). Moreover, in both our study and the recent study by Tang et al. (13), reporting a 12-gene signature predictive of adjuvant chemotherapy response, the identified gene signatures were not prognostic in squamous cell carcinomas, which comprise a large part of both the untreated and treated groups in the GSE14814 and GSE42127 cohorts. Furthermore, squamous cell carcinomas generally displayed higher expression of the CIN70 metagene than adenocarcinomas in all NSCLC cohorts included in the current study, leading to enrichment of these tumors in high proliferative risk groups when analyzed together with adenocarcinomas. Together, this highlights the need for adequately sized studies to identify and/or evaluate the clinical value of predictive gene signatures in a histology specific setting.

EGFR tyrosine kinase and ALK inhibitors have improved clinical outcome in advanced NSCLC (5, 45), although only a smaller fraction of tumors harbor EGFR mutations or ALK rearrangements and thus fulfill criteria for such treatment.
Moreover, these patients usually relapse because of primary or acquired resistance. Clearly, additional tools are needed to further guide diagnosis and treatment with targeted therapies in lung cancer patients. Response to EGFR inhibitors has been associated with retention of an epithelial phenotype in NSCLC cell lines and tumors by epithelial-mesenchymal transition (EMT) gene signatures derived from cell line experiments (26, 46). However, the performance of these EMT signatures in resected tumor tissue remains to be clarified in larger patient cohorts. We found only a limited overlap (5-6 genes) between the two multicohort signatures and one such EMT gene signature (26). Wilkerson et al. reported that bronchioid classified EGFR wild type tumors displayed higher average gefinitib sensitivity scores, based on a cell line expression signature, than non-bronchioid EGFR wild type tumors (19). This suggests that EGFR wild type tumors responding to gefinitib would be of the bronchioid subtype (19). Given that KRAS-mutant adenocarcinomas are resistant to EGFR inhibitors (5), this subset of tumors would then primarily correspond to our low-risk wt/wt-2 subgroup. By contrast, no significant difference in gefinitib sensitivity scores was observed between EGFR-mutated tumors stratified by molecular subtype (19). Yuan et al. recently reported that clustered genomic alterations (copy number gains) on chromosome 7p predicted clinical outcome and response to EGFR inhibitors in EGFR-mutated, but not in EGFR wild-type adenocarcinomas (47). Consistent with Yuan et al. (47), the high-risk EGFR-2 group showed more copy number alterations and more amplifications on chromosome 7p than the low-risk EGFR-1 group in EGFR-mutated tumors. Moreover, one of the representative genes from Yuan et al., VOPPI, is included in the EGFR-1/2 multicohort classifier with highest expression in EGFR-2. Additionally, EGFR itself was significantly upregulated in EGFR-mutated EGFR-2 tumors in two of the three cohorts defining the EGFR-1/2 classifier.
compared with EGFR-1 tumors, potentially due to a higher frequency of EGFR copy number gain or amplification in this subgroup. To further explore the association of our derived multicohort signatures with response to targeted treatment, we classified tumor biopsies from EGFRwt/KRASwt NSCLC patients with advanced chemorefractory metastatic disease treated with sorafenib enrolled in the BATTLE trial (26). We show that the high-risk EGFR-2 and the wt/wt-1 groups included more cases without disease control, and the wt/wt-1 group was borderline non-significant for worse progression-free survival. These results appear consistent with a more aggressive phenotype for the EGFR-2 and wt/wt-1 high-risk groups, as suggested by their overall higher CIN70 metagene expression, higher number of copy number alterations/amplifications, and association with a poorly differentiated tumor phenotype. In addition, our results, combined with previous reports (26, 48, 49), demonstrate the potential of applying prognostic/predictive gene expression signatures to small biopsy specimens from patients with non-operable disease, provided that enough tissue material could be sampled.

Together, these findings suggest that a connection between the identified transcriptional subgroups and response to different targeted treatments is possible, and that mutation status and molecular subtyping together could potentially predict therapy response better than mutation status alone. Patients with EGFR-mutated adenocarcinomas treated with targeted tyrosine kinase inhibitors represent the most important such case. In addition, the growing number of detected tyrosine kinase fusions (including ALK, RET, and ROS1) in predominantly lung adenocarcinoma are also becoming increasingly important as these alterations are/may become targets for specialized molecular agents and comprise a notable fraction of EGFRwt/KRASwt adenocarcinomas. With the exception of 11 confirmed ALK-positive cases in
GSE31210, patient specific information about ALK, RET, and ROS1 rearrangements were not available for included cohorts in the current study. Marked overexpression of these genes, which could indicate the presence of potential rearrangements, were only observed in a small number of cases across the different cohorts (data not shown). This was especially evident for ALK when compared with the expression levels of the known ALK-positive cases in GSE31210. Together, this precluded a detailed analysis of the multicohort signatures in these subgroups.

In summary, we identified transcriptional subgroups in EGFR-mutated and EGFRwt/KRASwt adenocarcinomas with clinical and genomic differences based on a multicohort discovery and validation strategy. The identified gene signatures also added independent prognostic information in a general lung adenocarcinoma context irrespective of mutation status, and showed promising associations with response to different treatments. Further analyses in larger well-characterized cohorts with available treatment response data for EGFR inhibitors or other therapeutic agents are required to determine the predictive values of the identified gene signatures in a mutation specific and general context.

**List of abbreviations used**

ACT: Adjuvant chemotherapy

EMT: epithelial-mesenchymal transition

EGFRwt/KRASwt: wild type for EGFR and KRAS

NSCLC: Non-small cell lung cancer

**Grant support**
Financial support for this study was provided by the Swedish Cancer Society, the Knut & Alice Wallenberg Foundation, the Foundation for Strategic Research through the Lund Centre for Translational Cancer Research (CREATE Health), the Mrs Berta Kamprad Foundation, the Gunnar Nilsson Cancer Foundation, the Swedish Research Council, the Lund University Hospital Research Funds, the Gustav V:s Jubilee Foundation, and the IngaBritt and Arne Lundberg Foundation.

Acknowledgement

The authors would like to thank the editors at Elevate Scientific for helpful comments on the manuscript.

References


Table 1. Clinical characteristics of adenocarcinoma patients in the prognostic gene expression cohorts.

<table>
<thead>
<tr>
<th></th>
<th>Discovery cohorts</th>
<th>Validation cohorts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GSE31210 (16)</td>
<td>GSE13213 (15)</td>
</tr>
<tr>
<td></td>
<td>Chitale U133A (22)</td>
<td>GSE42127 (13)</td>
</tr>
<tr>
<td></td>
<td>Chitale U133 2plus (22)</td>
<td>GSE3141 (24)</td>
</tr>
<tr>
<td></td>
<td>E-MTAB-923 (21)</td>
<td>Shedden (14)</td>
</tr>
<tr>
<td>Total no. of patients</td>
<td>226</td>
<td>117</td>
</tr>
<tr>
<td></td>
<td>91</td>
<td>133</td>
</tr>
<tr>
<td></td>
<td>102</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>103</td>
<td>356</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>105</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>41</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>42</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>189</td>
</tr>
<tr>
<td>Female</td>
<td>121</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>87</td>
<td>166</td>
</tr>
<tr>
<td>Smoking status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never-smokers</td>
<td>115</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>63</td>
<td>33</td>
</tr>
<tr>
<td>Smokers</td>
<td>111</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>73</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>83</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>229</td>
</tr>
<tr>
<td>Mutation status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EGFR-mutated</td>
<td>127</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>49</td>
<td>-</td>
</tr>
<tr>
<td>KRAS-mutated</td>
<td>20</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>36</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>-</td>
</tr>
<tr>
<td>EGFRwt/KRASwt</td>
<td>79</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>65</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>42</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>33</td>
<td>-</td>
</tr>
<tr>
<td>Stage</td>
<td>1</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>168</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>53</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>70</td>
<td>224</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td></td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>58</td>
</tr>
<tr>
<td>----</td>
<td>------</td>
<td>-----</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>IV</td>
<td>0</td>
</tr>
<tr>
<td>ACT*</td>
<td>226/0</td>
<td>Unknown</td>
</tr>
<tr>
<td>Median follow-up (Years)</td>
<td>5</td>
<td>3.5</td>
</tr>
<tr>
<td>Platform</td>
<td>Affymetrix U133 2plus</td>
<td>Affymetrix U133A</td>
</tr>
</tbody>
</table>

*: Adjuvant chemotherapy. Number of untreated / treated patients.
Figure legends

Figure 1. Schematic of the multicohort discovery and validation design.

Figure 2. Consensus clustering of discovery cohorts for \textit{EGFR}-mutated and \textit{EGFRwt/KRASwt} tumors. Consensus clustering was performed on normalized, mean-centered, and variance filtered gene expression data using k=2 groups as described (Supplementary Information). Heatmaps display consensus values between pairs of tumors by blue shading. High consensus corresponds to samples that always occur in the same cluster and is shaded dark blue. \textit{KRAS}-mutated tumors were omitted as no reproducible subgroups from clustering were found in later cross-cohort analyses. (A) Clustering of the four discovery cohorts for \textit{EGFR}-mutated adenocarcinomas. Note that the Chitale U133A cohort does not show distinct clusters, and later cross-cohort analyses could not reproduce subgroups from the other cohorts in this cohort. (B) Clustering of the four discovery cohorts for \textit{EGFRwt/KRASwt} adenocarcinomas.

Figure 3. Association with overall survival for \textit{EGFR}-mutated and \textit{EGFRwt/KRASwt} transcriptional subgroups.

(A) \textit{EGFR}-mutated subgroups (EGFR-1 and EGFR-2) display difference in overall survival. (B) \textit{EGFRwt/KRASwt} subgroups (wt/wt-1 and wt/wt-2) display difference in overall survival, with the exception of E-MTAB-923. (C) Multicohort \textit{EGFR}-1/2 centroid classification is associated with overall survival in \textit{EGFR}-mutated tumors in GSE13213. (D) Multicohort wt/wt-1/2 centroid classification is associated with overall survival in \textit{EGFRwt/KRASwt} tumors in GSE13213.
Figure 4. Multivariate Cox regression analyses of the EGFR-1/2 and wt/wt-1/2 signatures in independent adenocarcinoma cohorts.

Multivariate analyses of EGFR-1/2 and wt/wt-1/2 were performed using overall survival as endpoint. For GSE13213 tumor stage, patient age, mutation status, and smoking status were included as covariates in the multivariate analyses. For GSE42127 tumor stage, patient age, and adjuvant chemotherapy were included as covariates. For Shedden et al. tumor stage, patient age, adjuvant chemotherapy, and smoking status were included as covariates. Only hazard ratios and p-values for significant or borderline non-significant covariates are displayed. (A) EGFR-1/2 analysis for all cases for respective cohort. (B) wt/wt-1/2 analysis for all cases for respective cohort. (C) EGFR-1/2 analysis in stage I tumors for respective cohort. (D) wt/wt-1/2 analysis in stage I tumors for respective cohort.
Four discovery cohorts: GSE31210, Chitale U133A, Chitale U133 2plus, E-MTAB-923

EGFR-mutated  EGFRwt/KRASwt  KRAS-mutated

Unsupervised consensus clustering, SAM analysis, centroid cross cohort classification

Reproducible subgroups across multiple cohorts: EGFR-1/2 & wt/wt-1/2

Create multicohort nearest-centroids for EGFR-mutated and EGFRwt/KRASwt groups

One EGFR-1/2 multicohort & one wt/wt-1/2 multicohort classifier

Prognostic performance of each classifier in independent data

GSE13213  Shedden  GSE42127  GSE3141

Mutation specific  Overall Stage I Histology

Predictive performance of each classifier

GSE33072  GSE42127  GSE14814
Lung SPORE  JBR.10

Sorafenib in EGFRwt/KRASwt NSCLC  Adjuvant chemotherapy in NSCLC
Figure 3

A)

B)

C)

D)
Figure 4

A) Shedden
n=221: Stage III
GSE42127
n=132: EGFR−2
GSE13213
n=117: Stage III
GSE13213
n=117: EGFR−2

HR=1.04 (1.02-1.06)

p<0.00001
p= 5e-04
p= 0.03
p= 0.003
p= 0.002
p= 0.002
p= 0.005
p= 0.02

B) Shedden
n=220: Stage II
GSE42127
n=132: wt/wt−1
GSE13213
n=117: Stage III
GSE13213
n=117: wt/wt−1

HR=1.05 (1.02-1.07)

p<0.00001
p= 6e-05
p= 0.001
p= 0.005
p= 0.002
p= 0.004

C) Shedden Stage I
n=141: Age
GSE42127 Stage I
n=89: Age
GSE42127 Stage I
n=89: EGFR−2
GSE13213 Stage I
n=79: EGFR−2

HR=1.06 (1.02-1.09)

p= 7e-04
p= 0.01
p= 0.05
p= 0.02
p= 0.003

D) Shedden Stage I
n=141: Age
GSE42127 Stage I
n=89: Age
GSE42127 Stage I
n=89: wt/wt−1
GSE13213 Stage I
n=79: wt/wt−1

HR=1.06 (1.03-1.1)

p= 5e-04
p= 0.02
p= 0.06
p= 0.02
p= 0.01
Identification of transcriptional subgroups in EGFR-mutated and EGFR KRAS-wild type lung adenocarcinoma reveals gene signatures associated with patient outcome

Maria Planck, Sofi Isaksson, Srinivas Veerla, et al.

*Clin Cancer Res* Published OnlineFirst August 12, 2013.

<table>
<thead>
<tr>
<th>Updated version</th>
<th>Access the most recent version of this article at: doi:10.1158/1078-0432.CCR-13-0928</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supplementary Material</td>
<td>Access the most recent supplemental material at: <a href="http://clincancerres.aacrjournals.org/content/suppl/2013/08/12/1078-0432.CCR-13-0928.DC1">http://clincancerres.aacrjournals.org/content/suppl/2013/08/12/1078-0432.CCR-13-0928.DC1</a></td>
</tr>
<tr>
<td>Author Manuscript</td>
<td>Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.</td>
</tr>
</tbody>
</table>

E-mail alerts Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.