Genome-wide DNA methylation analysis of lung carcinoma reveals one neuorendocrine and four adenocarcinoma epitypes associated with patient outcome

Anna Karlsson¹, Mats Jönsson¹, Martin Lauss¹, Hans Brunnström¹, Per Jönsson², Åke Borg¹, Göran Jönsson¹,³, Markus Ringnér¹,³, Maria Planck¹, Johan Staaf¹,³,*

¹Division of Oncology and Pathology, Department of Clinical Sciences Lund, Lund University, Medicon Village, SE 22381 Lund, Sweden
²Department of Thoracic Surgery, Lund University and Skåne University Hospital, SE 22185 Lund, Sweden
³CREATE Health Strategic Center for Translational Cancer Research, Lund University, Medicon Village, SE 22381, Lund, Sweden
*Corresponding author

Running title: Epitypes of lung carcinoma

Key words: lung cancer, SCLC, LCNEC, adenocarcinoma, DNA methylation, CpG, mutation, copy number alteration, EGFR, KRAS, overall survival, gene expression

Corresponding author:
Johan Staaf, e-mail: johan.staaf@med.lu.se, Address: Division of Oncology and Pathology, Department of Clinical Sciences Lund, Lund University, Medicon Village, SE 22381 Lund, Sweden, Fax: +4646147327
Competing interests

Authors declare that they have no competing interests.

Nbr figures / tables: 6        Word count: 4981
Translational relevance

DNA methylation in gene promoters is a major mechanism of gene expression regulation that may promote tumorigenesis. DNA methylation of specific genes, and/or patterns of DNA methylation in lung cancers have been associated with patient outcome. However, hitherto, neither the existence of reproducible DNA methylation-based subgroups of potential clinical relevance nor the DNA methylation pattern across multiple histological subgroups has been carefully investigated in lung cancer. Based on a multicohort approach, we conducted a comprehensive survey of genome-wide DNA methylation patterns in lung cancer identifying one neuroendocrine and four reproducible adenocarcinoma epitypes. Importantly, epitypes were associated with specific clinicopathological and molecular characteristics, gene expression phenotypes, and patient outcome. These findings shed light on the epigenetic characteristics and molecular diversity underlying lung cancer. Moreover, they highlight the possibility to further subgroup the disease based on epigenetic/molecular classification, which could lead to improvements in tumor classification, prognostication, and tailored patient therapy.
Abstract

Purpose
Lung cancer is the worldwide leading cause of death from cancer. DNA methylation in gene promoter regions is a major mechanism of gene expression regulation that may promote tumorigenesis. However, whether clinically relevant subgroups based on DNA methylation patterns exist in lung cancer remains unclear.

Experimental Design
Whole-genome DNA methylation analysis using 450K Illumina BeadArrays was performed on 12 normal lung tissues and 124 tumors including 83 adenocarcinomas, 23 squamous cell carcinomas (SqCC), one adenosquamous cancer, five large cell carcinomas, nine large cell neuroendocrine carcinomas (LCNEC), and three small cell carcinomas (SCLC). Unsupervised bootstrap clustering was performed to identify DNA methylation subgroups, which were validated in 695 adenocarcinomas and 122 SqCCs. Subgroups were characterized by clinicopathological factors, whole-exome sequencing data, and gene expression profiles.

Results
Unsupervised analysis identified five DNA methylation subgroups (epitypes). One epitype was distinctly associated with neuroendocrine tumors (LCNEC and SCLC). For adenocarcinoma, remaining four epitypes were associated with unsupervised and supervised gene expression phenotypes, and differences in molecular features including global hypomethylation, promoter hypermethylation, genomic instability, expression of proliferation-associated genes, and mutations in KRAS, TP53, KEAP1,
SMARCA4, and STK11. Furthermore, these epitypes were associated with clinicopathological features such as smoking history, and patient outcome.

**Conclusions**

Our findings highlight one neuroendocrine and four adenocarcinoma epitypes associated with molecular and clinicopathological characteristics, including patient outcome. This study highlights the possibility to further subgroup lung cancer, and more specifically adenocarcinomas, based on epigenetic/molecular classification that could lead to more accurate tumor classification, prognostication, and tailored patient therapy.
Introduction

Lung cancer is currently the leading cause of death from cancer worldwide (1). The disease is broadly divided into small cell lung cancer (SCLC, ~15% of all cases) and non-small cell lung cancer (NSCLC). NSCLC is further divided into adenocarcinoma, squamous cell carcinoma (SqCC) and large cell carcinoma with or without neuroendocrine features (LCNEC and LC, respectively). Lung cancer is a molecularly heterogeneous disease involving different alterations that drive tumorigenesis, including DNA sequence alterations, copy number alterations (CNAs), and epigenetic modifications, including DNA methylation and histone/chromatin modifications. DNA methylation at CpG dinucleotides in gene promoter regions is a major mechanism of gene expression regulation, and aberrant promoter hypermethylation may lead to inactivation of tumor suppressor genes, thereby promoting tumorigenesis (2).

Genome-wide DNA methylation profiling of NSCLC have identified epigenetic subgroups (epitypes) of tumors associated with characteristic molecular alterations and prognosis (3-7). NSCLC/adenocarcinomas with increased promoter methylation levels have been highlighted, and termed CpG Island Methylator Phenotype (CIMP) stratifying tumors into CIMP-high, CIMP-low/negative and CIMP-intermediate subgroups, in analogy to findings from other cancer forms (3, 4, 8). Additionally, gene expression phenotypes like the bronchioid, magnoid, and squamoid subtypes in adenocarcinoma (9, 10), have also been associated with specific DNA methylation patterns (9). However, the proposed NSCLC epitypes have not been independently replicated. Moreover, genome-wide epigenetic patterns across multiple lung cancer histotypes have not yet been reported.
Herein, we investigated the landscape of DNA methylation in different histological subgroups of lung cancer with the intention to derive methylation-based subgroups of clinical and molecular relevance. Based on a discovery cohort of 124 primary lung cancers, including all major histological subgroups, we found a specific DNA methylation pattern of neuroendocrine tumors and identified four epitypes of adenocarcinoma that were subsequently validated in 817 independent NSCLC cases. Epitypes were associated with molecular and clinicopathological differences, and linked to gene expression phenotypes based on integration with DNA sequencing and gene expression data. Together, our findings highlight the possibility to further subgroup lung cancer based on epigenetic/molecular classification, providing a clear refinement of previously suggested models and a more accurate tumor classification, which could lead to new targets for diagnostics, therapeutic intervention, and prognostication of the disease.

Materials and Methods

Patient material
DNA and total RNA were extracted from the same tissue piece for 124 tumors and 12 matched normal lung tissue specimens from early stage lung cancer patients operated at the Skåne University Hospital in Lund, Sweden (discovery cohort, Table 1). The study was approved by the Regional Ethical Review Board in Lund, Sweden (Registration no. 2004/762 and 2008/702). The 12 normal specimens originated from patients with adenocarcinoma and were mixed in gender (three males, three females), smoking status (six never-smokers, six smokers) and patient age (57-82 years). 373 adenocarcinomas from The Cancer Genome Atlas (TCGA) project (11) and 444
NSCLC cases from Sandoval et al. (5) (Sandoval) were used as validation cohorts (Table 1).

**Global methylation analysis**

All cases were analyzed by the Illumina Human Methylation 450K v1.0 platform (Illumina, San Diego, CA, US) according to manufacturer’s instructions (Supplementary Methods). Signal intensities were obtained from GenomeStudio (Illumina), converted to beta-values, filtered, and normalized to remove biases between Infinium I and II probes (Supplementary Methods). CpG probes with aberrant methylation in tumors compared to normal lung tissue in the discovery set were identified as described in Supplementary Methods (Figure S1A), and annotated through the human embryonic stem cell (H1hESC) chromatin state track (12) and the Illumina CpG island track. CpGs in repetitive elements were identified through the “repeats_rmsk_hg19” table from the UCSC Genome Browser. Unsupervised class-discovery was performed using bootstrap clustering (13) (Supplementary Methods). Principal component analysis (14), including clinicopathological and technical factors, and comparison of bisulfite conversion plate and beadchip id against unsupervised bootstrap clusters were performed to assess that no technical artifacts influenced methylation data, or bootstrap groups, for the 124-sample discovery cohort (Figures S1B-D). DNA methylation centroids representing bootstrap clusters were created from the average beta-value for each CpG probe in respective cluster. Samples in validation cohorts were assigned to the centroid with the smallest Euclidean distance for matching CpGs. Methylation data for the discovery cohort is available as GSE60645 (15).
Copy number analysis

Log2 copy number estimates and CNAs for CpG probes in the discovery and Sandoval cohorts were generated and identified as described in Supplementary Methods from 450K methylation beadchip data. For the TCGA cohort copy number estimates and CNAs were obtained from level 3 Affymetrix SNP6 data as described (16, 17). Complex arm-wise aberration index (CAAI) scores were calculated similar to Russnes et al. (18) (Supplementary Methods).

Global gene expression analysis

Gene expression analysis was performed on 117 tumors from the discovery cohort using Illumina Human HT-12 V4 microarrays, available as GSE60645 (15). TCGA adenocarcinoma expression data were obtained as RNASeq V2 data. Six correlated gene expression modules in lung cancer, representing different tumor and/or tumor environment associated processes, were derived as originally described by Fredlund et al. in GSE29016 (19, 20) (Supplementary Methods and Table S1). These expression modules included an immune response, a neuroendocrine, and a stroma/extra cellular matrix module. Data processing steps, including adenocarcinoma and SqCC molecular subtype classification (9, 21), correlation of methylation and expression data, and calculation of different expression metagenes are further described in Supplementary Methods.

Functional classification

Gene Ontology enrichment were performed using the DAVID Functional Annotation Tool (22) with the default human population background and a Bonferroni-adjusted p-value <0.05 as significance threshold.
Results

Genome-wide DNA methylation patterns in lung cancer

We analyzed 124 lung tumors from five histological subgroups for global DNA methylation patterns using Illumina 450K methylation arrays (Table 1, discovery cohort). Overall, DNA methylation in the tumors followed a distinct pattern along the gene coding sequence, with low methylation levels near the transcription start site and high methylation levels at gene bodies, 3'UTRs, and intergenic regions (Figure 1A). Correlation analyses of DNA methylation and gene expression revealed a pattern of negative correlations at transcription start sites and more positive correlations in open sea/heterochromatin regions and gene bodies (Figures 1B and C, Table S2). We identified 4136 CpGs with aberrant methylation in >10% (n=13) of tumors compared to normal lung tissue, including multiple HOX genes, Wnt signaling pathway genes, APC, CDH13, GATA4, GATA5, and RASSF1 consistent with previous studies (3, 6) (Figure S1A, Table S2). Hypomethylated CpGs in tumors were enriched in open sea/heterochromatin regions, whereas hypermethylated CpGs were typically located in transcription start sites, CpG islands, and poised promoters in human embryonic stem cells, H1hESC cells, consistent with previous reports (6, 23) (Figures 1D and E). Hypomethylated CpGs were enriched in repetitive regions (LINE, SINE, LTR elements) compared to hypermethylated CpGs (21% vs. 4%, respectively, Fisher’s exact p=9e-54). Importantly, changing the number of CpGs with aberrant methylation by lowering or increasing the number of required tumors with aberrant methylation...
(n=2-20 tumors equaling CpG-sets between ~1000-44000 CpGs, Figure S1A) yielded the same enrichment pattern of hypo- and hypermethylated CpGs.

Functional annotation analysis of genes with hypermethylated CpGs in the 4136 CpG-set showed enrichment of biological processes such as regulation of transcription, neural development, and cell morphogenesis corroborating previous studies (24, 25), while hypomethylated genes showed a much less clear functional enrichment (Table S2).

*Unsupervised class discovery based on genome-wide DNA methylation patterns identifies five epitypes*

Unsupervised bootstrap analysis based on the 4136 CpGs highlighted five tumor clusters in the discovery cohort, hereafter referred to as epitypes (ES1, ES2, ES3, ES4, and ES5) (Figures 2A and S1E). Importantly, epitype association for individual samples was robust across different CpG sets (numbers between 1282-17710 CpGs) in exploratory bootstrap analysis (Figure S1F). ES1 showed a global hypomethylation pattern, ES4 a promoter methylation pattern, ES5 a methylation pattern resembling normal lung tissue, whereas ES2 had a pattern in between ES1 and ES4 (Figure 2). Consistent with the global DNA hypomethylation pattern, ES1 tumors also showed more hypomethylation of CpGs in repetitive elements (Figures 2C and S1G). Notably, 89% of ES3 cases were either SCLC (n=2) or LCNEC (n=6) tumors. Consistent with the dominance of neuroendocrine cases in ES3 we found distinct overexpression of a neuroendocrine gene expression metagene compared to the other epitypes (p=5e-05, Kruskal-Wallis test). Hence, we refer to ES3 as a neuroendocrine epitype. On the other hand, SqCC tumors clustered in ES1 (17%), ES2 (57%), and ES5 (22%) (Figure 2A). A distinct association of SqCC cases in ES2 with the reported classical SqCC
gene expression subtype (21) was found, with >86% of classical subtype classified SqCC cases present in this epitype. Adenocarcinomas (n=83) were divided into ES1 (12%), ES2 (14%), ES4 (36%) and ES5 (37%) (Figure 3A).

Validation of lung cancer epitypes

To validate the identified epitypes from the discovery cohort we created DNA methylation centroids for each epitype based on the 4136 CpGs. Next, we classified two independent cohorts analyzed by the same methylation platform (Sandoval and TCGA) comprising 122 SqCC tumors and 695 adenocarcinomas (Table 1). Principal component analysis performed in the validation cohorts confirmed that the centroid classification explained most of the total variation in DNA methylation compared to available clinicopathological, technical (batch and beadchip data) and molecular factors, including clinical smoking history, sex, tumor stage, tumor size, histology (adenocarcinoma or SqCC), EGFR, KRAS, and TP53 mutations (Figures S2A, S3A-C). Notably, most of these factors (e.g. smoking status) contributed little to the total variation in DNA methylation. Moreover, the classification of the validation cohorts was robust across different sets of CpGs, and overlapped extensively with independently derived unsupervised bootstrap groups in these cohorts (Figures S2B-D and S3D-F).

In both validation cohorts, ≤1% of cases were classified as ES3, supporting that this epitype is highly distinct for lung cancers expressing neuroendocrine marker genes. Similar to the discovery cohort, SqCC tumors in the Sandoval cohort were primarily classified as ES2 (49% of SqCC cases) or ES5 (33%). Although LC, LCNEC, and SqCC tumors were present in different clusters in the discovery set this cohort is underpowered to robustly claim existence of different epitypes within these
histological subgroups. Moreover, there currently exist no comparable LC, LCNEC, or SCLC cohorts suitable for validation of novel epitypes within these subgroups. Consequently, we hereafter focus the characterization and validation of the epitypes only on lung adenocarcinomas in the three cohorts (excluding ES3), using clinicopathological factors, gene expression data, CNAs, and mutational data. Figure 3A shows the distribution of adenocarcinomas between epitypes in all investigated cohorts.

Adenocarcinoma epitypes are associated with reproducible clinicopathological characteristics including smoking history, EGFR and KRAS mutations

The epitypes showed differences in the composition of never-smokers and smokers. ES5 was enriched for never-smokers in both the discovery and Sandoval cohorts (63-68% of all never-smokers), while less in the TCGA cohort (35%) (Figure 3B). By contrast, never-smokers were rarely classified as ES1 in any cohort (0-5% of all never-smokers). However, in exploratory analysis we identified only 513 CpGs (1.1% of analyzed CpGs) to be statistically associated with clinical smoking status in adenocarcinomas across all three cohorts (false discovery rate adjusted Wilcoxon p<0.05 and >0.05 difference in average beta-value between groups). Notably, only 21 of these CpGs showed a more stringent difference in DNA methylation (>0.1 average beta-value difference).

Consistent with the distribution of never-smokers, EGFR mutations were often found in ES5 tumors in the discovery and TCGA cohorts (58% and 30% of all mutations, respectively), but rarely in ES1 cases (4-8%) (Figure 3C). Another notable difference between the epitypes was a similar enrichment of KRAS-mutated cases in
the ES4 promoter hypermethylated epitype in both the discovery and TCGA cohorts (50-54% of all KRAS mutations) (Figure 3C).

*Adenocarcinoma epitypes are associated with adenocarcinoma gene expression phenotypes*

In both the discovery and TCGA cohorts, the epitypes were associated with the reported bronchioid (ES5), magnoid (ES1, ES2) and squamoid (ES4) adenocarcinoma gene expression phenotypes (9) (Figure 3D). The association of the epitypes with gene expression phenotypes was further supported by an extensive overlap between epitypes and gene expression subgroups derived from individual unsupervised consensus clustering of the discovery and TCGA cohorts (Figure 3E). Together, these results provide a strong link between genome-wide DNA methylation and the transcriptional landscapes in lung adenocarcinoma.

*Gene expression signatures associated with adenocarcinoma epitypes*

The epitypes were associated with consistent differences in various gene expression metagenes in both the discovery and TCGA cohorts. For instance, ES1 had the highest expression of proliferation-associated genes (the CIN70 (26) metagene), while ES5 the lowest (p=0.00005 in the discovery cohort and p=9e-17 in TCGA, Kruskal-Wallis test). The opposite pattern was found for expression of a terminal respiratory unit (TRU) (27) gene signature (p=0.0002 and p=6e-18, respectively, Kruskal-Wallis test). The epitypes also differed in expression of an immune response associated metagene and a stroma/extracellular matrix associated metagene. Notably, the expression of these two gene modules likely relates to infiltration of immune or stromal cells in the analyzed macrodissected tissue. ES1 consistently showed the
lowest and ES5 the highest expression of both metagenes (Figure 3F, data not shown for the TCGA cohort). These results suggest that ES5 is an epitype with considerable infiltration of non-malignant cells consistent with the observed methylation pattern being most similar to normal lung tissue. By contrast, ES1 would represent tumors with high tumor cell content. ES2 showed a different pattern for these two metagenes compared to the other epitypes (Figure 3F). Expression of the stromal metagene was similar in ES2, ES4, and ES5, whereas expression of the immune metagene was lower in ES2 compared to ES4 and ES5, but higher compared to ES1. Supporting these observations, we found similar patterns of stromal and immune expression scores between the epitypes using the Estimation of Stromal and Immune Cells in Malignant Tumors (ESTIMATE) method (28) in both cohorts (data not shown). Together, this suggests that differences in the cellular type and amount of infiltrating non-malignant cells may exist between the epitypes.

To further investigate biological processes differing between the epitypes we identified differentially expressed genes between adenocarcinomas stratified by epitype in the discovery (n=1824 expression probes) and TCGA (n=5726 genes) cohorts (Supplementary Methods). Functional analysis revealed enrichment of biological processes involved in immune response, cell proliferation and cell adhesion (Supplementary Table S3), consistent with results from the metagene analyses (Figure 3).

The mutational spectrum of adenocarcinoma epitypes

To further characterize the mutational spectrum in the epitypes we analyzed whole-exome sequencing data for TCGA adenocarcinomas. Overall, ES1 cases harbored the highest number of mutations and ES5 the least (Figure 4A), independent of patient
smoking status. Moreover, the epitypes showed differences in the type of mutation transversions when stratified by smoking status (Figure S4). The largest differences were observed in the distributions of C>T and C>A transversions (recognized as a smoking-related signature (29)), between the ES1 (more C>A, less C>T) and ES5 epitypes (less C>A, more C>T). Consistently, overlapping ES1 cases were more often classified as transversion-high (89%) in the recent TCGA study compared to the other epitypes (55-70%, Fisher’s exact p=0.03) (3).

To search for individual mutations associated with the epitypes we performed a permutation-based screen of 174 genes identified by MutSigCV (30) analysis of 402 TCGA adenocarcinomas as described (31). This analysis identified seven genes with false discovery rate ≤10%, including four well-known tumor suppressor genes (KEAP1, TP53, STK11, and SMARCA4) and three genes appearing as either false positives (COL11A1, and LRRIQ3), or with <10% mutation frequency in any epitype (SNRPN). For TP53, STK11, KEAP1, and SMARCA4, we observed notable differences in the mutation frequencies between the epitypes (Figure 4B), but no differences in mutation type (missense, truncating, or in-frame indel, Chi-square p>0.05). The latter result may partly be related to the overall low number of specific mutations, as e.g. 86% of SMARCA4 mutations in ES4 were missense mutations compared to 30-50% in ES1, ES2, or ES5. In these analyses, KRAS mutations were borderline non-significantly associated with the epitypes, whereas the association of EGFR mutations with the epitypes was less strong (see Figure 3C).

**Adenocarcinoma epitypes are associated with patient outcome**

The four epitypes were associated with patient outcome (overall survival or relapse-free survival) in the discovery and Sandoval cohorts for NSCLC in general, and
adenocarcinoma specifically (Figure 5). Convincingly, in both cohorts, the ES2 and ES5 epitypes were associated with the best outcome in adenocarcinomas, while ES1 and ES4 with the worst. For stage I adenocarcinomas, the epitypes were associated with overall survival in the discovery cohort (log-rank p=0.005), while borderline non-significant in the Sandoval cohort (log-rank p=0.06). However, for NSCLC stage I tumors from Sandoval the epitypes were associated with relapse-free survival (log-rank p=0.04).

In univariate analysis of epitype association, patient age, smoking history, sex, EGFR and KRAS mutation status in stage I adenocarcinomas from the discovery cohort, the epitypes was the only significant factor for overall survival (p<0.05). In multivariate analysis including all these factors, the ES2 and ES5 epitypes remained significant (p<0.05). In multivariate analysis of stage I adenocarcinomas from the Sandoval cohort none of the factors (age, smoking history, gender, and epitype) reached significance.

**Discussion**

In the current study, we have explored the landscape of genome-wide DNA methylation across the major histological subgroups of lung cancer, identifying five epitypes of tumors linked to different gene expression phenotypes. We demonstrate that aberrant DNA methylation in lung cancer is consistent with the classical view of hypermethylation in CpG islands, and hypomethylation in heterochromatin regions, including repetitive elements (32). Hypermethylated genes were enriched for developmental and differentiation-associated processes and polycomb targets premarked by histone H3K27 tri-methylation in embryonic cells (24, 25). These results are consistent with a hypothesis that DNA methylation in lung cancer
preferentially targets genes involved in morphogenetic processes and late stage differentiation of the lung epithelium, potentially contributing to establishment of an early undifferentiated cancer phenotype (24).

Through a multicohort approach, we demonstrate that LCNEC and SCLC tumors with neuroendocrine features represent a distinct lung cancer epitype compared to NSCLC, consistent with a similar association based on copy number and transcriptional alterations (33). Supporting ES3 as a distinct neuroendocrine epitype, centroid classification of 69 NSCLC cell lines (7) classified only the known LCNEC cell line, NCI-H1155, as ES3. Remaining cell lines were predominantly classified as ES1 (58%) or ES4 (36%). In both the discovery and Sandoval NSCLC validation cohorts, DNA methylation epitypes identified by unsupervised bootstrap analysis comprised of a mix of adenocarcinomas and SqCCs. On the transcriptional and CNA level, adenocarcinomas and SqCCs display large differences (16, 27, 33). Here, additional studies (larger cohorts) are needed to pinpoint DNA methylation alterations that could explain such histology or cell type specific expression patterns.

In the discovery cohort, we divided adenocarcinomas (91% stage I tumors) into four epitypes (ES1, ES2, ES4, and ES5), with marked differences in molecular and clinicopathological characteristics, including patient outcome. Although resected stage I NSCLC patients have the most favorable prognosis, the 5-year survival rate is 52-89% (34). Thus, improved molecular subclassification of early stage NSCLC is highly relevant. To date, only a few studies have reported DNA methylation epitypes in NSCLC or adenocarcinoma specifically (3-7). However, thorough validation of the reported epitypes has not been performed in any of these studies. In contrast, we validated our epitypes in 695 adenocarcinomas from two independent cohorts showing that they: i) provide powerful explanations of the total variation in DNA
methylation compared to other clinicopathological and molecular factors, ii) are robust across a wide range of CpGs, iii) have consistent clinicopathological and molecular features in different cohorts, and iv) could be recovered in validation cohorts by independent unsupervised analysis.

Based on extensive promoter hypermethylation, overrepresentation of KRAS-mutated adenocarcinomas, and poor patient outcome, the ES4 epitype shares features with the Shinjo et al. (4) adenocarcinoma CIMP-high phenotype. Supporting this association, 89% of all matching CIMP-high cases in the recent TCGA study were classified as ES4, while the remaining 11% were classified as ES1 (3). However, in contrast to the Shinjo et al. (4) CIMP-high phenotype the ES4 epitype included never-smokers, EGFR mutations, and was not associated with gender (similar to the CIMP-high group in (3), Figure 3). These discrepancies may be due to that the CIMP definition in lung cancer is not standardized, illustrated by the differences in CIMP-high frequency between the TCGA and Shinjo et al. studies (20.4% and 7.8%, respectively) (3, 4). Notably, the enrichment of KRAS-mutated adenocarcinomas in a promoter hypermethylated cluster is consistent with previous reports (4, 6). This enrichment is intriguing given that KRAS-mutated adenocarcinomas have been reported to display less distinctive mRNA and CNA patterns compared to, e.g., EGFR-mutated adenocarcinomas (17, 35). However, KRAS mutations have not been found to be the driver of such a promoter hypermethylated epitype in either lung adenocarcinoma or colorectal cancer, suggesting a more complex underlying mechanism (6, 36).

The ES1 epitype was characterized by global hypomethylation distant from CpG islands, hypomethylation of CpGs in repetitive elements, high expression of proliferation-associated genes, a non-TRU-like expression pattern, association with
the magnoid subtype, a high mutational burden including TP53, KEAP1 and STK11 mutations, strong association with smoking, and poor patient outcome in both discovery and validation cohorts (Figures 2-5). Hypomethylation in cancer have been associated with different repetitive elements that could contribute toward genomic instability (37, 38) (and references therein). Accordingly, we found that ES1 tumors displayed not only more CNAs, but also that these alterations appeared more complex compared to the other epitypes based on the complex arm-wise aberration index (CAAI) (18) (Supplementary Figure S5A-C). We also found that copy number breakpoints occurring in repetitive elements for copy number gain or loss regions were hypomethylated to a greater extent in ES1 tumors (Supplementary Figure S5D-E). Together, these clinicopathological and molecular characteristics suggest that tumor progression in ES1 may be primarily driven by genomic instability and less by classical oncogene activation (exemplified by a lower EGFR and KRAS mutation frequency). The latter is supported by that ES1 cases were less often denoted oncogene-positive compared to tumors from other epitypes based on data from the recent TCGA study (p=0.02, Fisher’s exact test) (3). Moreover, from the same study ES1 cases showed higher tumor purity and tumor ploidy compared to the other epitypes (Kruskal-Wallis p<0.0001). Thus, ES1 appears to represent a poorly differentiated, aneuploid, and aggressive subset of adenocarcinomas with high tumor cellularity, less driven by oncogene activation. A smaller fraction of ES1 cases showed concomitant global hypomethylation and promoter methylation (more evident in the larger validation cohorts, Figures S2E and S3G). This subset of cases may better resemble the Shinjo et al. (4) CIMP-high group, as they were all smokers and did not harbor any EGFR mutations (data not shown). Irrespectively, our results
support that a CIMP phenotype can occur in adenocarcinomas with markedly
different epigenetic, transcriptional, and genetic make-up.

DNA methylation patterns may act as a fingerprint for different cell types
(39). Compared to the DNA methylation pattern of ES1, the ES2 epitype appears
more infiltrated by non-malignant cells. Consistently, we observed differences in gene
expression of metagenes associated with immune response and stroma/extracellular
matrix between ES1 and ES2. Intriguingly, despite indicators of poor prognosis,
including frequent CNAs, higher expression of proliferation-related genes, association
with the magnoid expression subtype, and a high mutational burden (including TP53,
STK11, and KEAP1 mutations), ES2 adenocarcinoma cases (together with ES5 cases)
showed the best outcome. While the generally better prognosis of ES5 cases may be
attributable to their lower proliferation rate, the better prognosis of ES2 patients
compared to ES1 could, hypothetically, be related to an altered and/or reduced
immune cell infiltration in ES1, which have been shown to confer a poorer prognosis
in multiple cancer types (19, 40-42). Whether the ES2 epitype represents an
intermediate/transition state to ES1 for adenocarcinomas remains to be investigated.
Although somatic alterations in specific epigenetic regulators were recently found in a
notable proportion of adenocarcinomas, there were no associations with global DNA
methylation patterns (3). Here, the association of SMARCA4 (a nucleosome
remodeller) mutations with ES2 is intriguing and warrants further investigation.

In contrast to the other epitypes, ES5 showed a DNA methylation pattern with
similarities to blood leukocytes and normal lung tissue. Together with its more TRU-
like expression pattern, lower expression of proliferation-related genes, higher
expression of immune and stroma-related metagenes, high frequency of bronchioid
classified tumors, enrichment of never-smokers, and better patient outcome, ES5
matches a proposed TRU type of adenocarcinoma (43) but also shares characteristics with the CIMP-negative epitype reported by Shinjo et al. (4). Importantly, ES5 cannot be dismissed as an epitype merely due to sampling issues, as the analyzed tumor DNA carried both CNAs and mutations. For instance, for the 25 cancer hallmark genes defined by Imielinski et al. (44), 78% of ES5 cases in the TCGA cohort carried at least one alteration (mutation or CNA). Moreover, the lack of NSCLC cell lines classified as ES5 or ES2 (see above) does not dismiss these epitypes in clinical tumor specimens as for instance the well established intrinsic molecular subtypes in breast cancer are not reproduced exactly in breast cancer cell lines (45).

DNA methylation of smaller sets of CpGs / genes (e.g. CDKN2A, FHIT, APC, and RASSF1A) have been associated with smoking in both genome-wide and gene-focused studies of lung cancer (6, 38, 46) (and references therein). However, on a genome-wide level our results suggest an overall less dominant effect of clinical smoking history on the DNA methylation landscape in primary adenocarcinomas, despite the enrichment of never-smokers in specific epitypes. This conclusion may be exemplified by results from the principal component analyses, the presence of never-smokers in all epitypes with exception of ES1, the low number of smoking associated CpGs in both the current and previous studies (6, 46), and that the patterns of global and promoter hypermethylation between epitypes were similar irrespective of smoking status (Figures 2A, S1B, S2A, S3A and S6A-B). Combined with similar findings of intrinsically heterogeneous gene expression and CNA patterns in smoking-defined adenocarcinoma subgroups (20, 31), our results question if never-smokers can be identified as a molecular subgroup of its own with transcriptional, DNA methylation and CNA patterns clearly different from tumors arising in smokers. Instead, our study further supports that a majority of adenocarcinomas arising in
never-smokers together with a specific subset of tumors from smokers represent a more distinct and relevant molecular/biological entity of less aggressive and potentially more smoking-unrelated disease (20, 31). The clinical smoking definitions are intrinsically problematic due to their self-reported nature, but also because they do not capture the intensity and duration of cigarette exposure, and the exposure to environmental tobacco smoke and other pollutants for never-smokers. Interestingly, the few TCGA never-smokers classified as ES1 display smoking characteristic C>A transversion frequencies similar to current-smokers, clearly different from, e.g., ES2 classified never-smokers (Figure S4A). Thus, whether these never-smokers are “true” never-smokers remains unclear. This suggests that ES1 is in fact strongly related to patients with a smoking history and, importantly, presumably also distinct underlying tumor biology and/or tumorigenic events.

The question of whether the observed DNA methylation epitypes / alterations are driver or passenger events, and their position and role in the evolutionary tree of a tumor, remains to be determined. Promoter hypermethylation of individual genes, notably tumor suppressors like \textit{CDKN2A}, have been recognized as early events in lung tumorigenesis, while there is a lack of consensus over whether global hypomethylation is an early or late event in lung cancer (see (37, 38)). The impact of smoking on epigenetic modifications may further complicate the picture, as certain alterations have been associated with duration or amount of tobacco smoking and may thus be later events in the cancer development and progression (38). Whole genome bisulfite sequencing combined with other profiling/sequencing techniques, may be one potential way of reconstructing the evolution of a tumor in relation to driver mutations, CNAs and DNA methylation, as recently described for DNA alterations in breast cancer (47).
Besides describing DNA methylation patterns in lung adenocarcinoma, our study strongly supports a link between adenocarcinoma gene expression phenotypes and genome-wide DNA methylation patterns (9). Importantly, this link brings further insights and explanation to the observed clinicopathological characteristics, gene expression patterns, mutational signatures, and biological pathways/processes associated with the epitypes (see e.g. (3, 9, 27, 43)). However, the current study also extends the knowledge about genome-wide DNA methylation patterns in the adenocarcinoma gene expression phenotypes, e.g., showing that the current definition of these phenotypes comprises of a mix of DNA methylation patterns (Figure 3D). In opposite to Wilkerson et al. (9), we found that the magnoid subtype was strongly associated with a global DNA hypomethylation pattern in both the discovery and TCGA cohorts (Figures S6C and D). Furthermore, DNA methylation patterns in and between the epitypes were consistent irrespective of bronchioid, magnoid, or squamoid classification (Figures S6E and F). Together, our results suggest that further refinement of both the proposed gene expression phenotypes and the CIMP phenotype in lung adenocarcinoma should be possible through integrated analysis of transcriptional, copy number and DNA methylation data.

Epigenetic alterations, including DNA methylation, are potentially reversible which offers an interesting therapeutic opportunity. For instance, DNA methyltransferase (DNMT) inhibitors can induce DNA hypomethylation at specific gene loci that can result in sustained gene reactivation (48). Currently, DNMT inhibitors and multiple histone deacetylase (HDAC) inhibitors are in clinical use and/or clinical testing in different malignancies, and a recent phase I/II trial reported an objective response to a combinatorial treatment with DNMT and HDAC inhibitors in recurrent metastatic NSCLC (49). Interestingly, in a recent NSCLC cell line
experiment, cell lines with a CIMP-positive phenotype responded with growth inhibition to 5-Aza-dC (a DNMT inhibitor) treatment, while CIMP-negative cell lines did not (4). Whether the proposed epitypes in the current study define patient subgroups likely to benefit or not from such treatment remains to be investigated.

In summary, based on a multicohort approach we have conducted a comprehensive survey of the genome-wide DNA methylation pattern in lung cancer involving the major histological subgroups. Together, the current study adds further layers of information about the epigenetic characteristics and molecular diversity in lung cancer. Moreover, it highlights the possibility to further refine disease classification that may ultimately lead to improvements in detection, patient stratification, prognostication, and therapy.

Acknowledgements

The authors thank the editors at Elevate Scientific for helpful comments on the article. 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. The SCIBLU Genomics center is supported by governmental funding of clinical research within the national health services (ALF) and by Lund University.

References

26
34. Crino L, Weder W, van Meerbeeck J, Felip E. Early stage and locally advanced (non-metastatic) non-small-cell lung cancer: ESMO Clinical Practice
# Tables

## Table 1. Patient characteristics and clinicopathological data for included cohorts.

<table>
<thead>
<tr>
<th></th>
<th>Lund cohort</th>
<th>Sandoval et al. (5)</th>
<th>TCGA (11)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Usage</strong></td>
<td>Discovery</td>
<td>Validation</td>
<td>Validation</td>
</tr>
<tr>
<td><strong>Total nbr patients</strong></td>
<td>124</td>
<td>444</td>
<td>373</td>
</tr>
<tr>
<td><strong>Histology</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adenocarcinoma</td>
<td>83</td>
<td>322</td>
<td>373</td>
</tr>
<tr>
<td>SqCC</td>
<td>23</td>
<td>122</td>
<td>-</td>
</tr>
<tr>
<td>LC</td>
<td>5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LCNEC</td>
<td>9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SCLC</td>
<td>3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Other</td>
<td>1 (Adenosquamous)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Tumor stage</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>110</td>
<td>237</td>
<td>206</td>
</tr>
<tr>
<td>II</td>
<td>10</td>
<td>94</td>
<td>87</td>
</tr>
<tr>
<td>III</td>
<td>1</td>
<td>102</td>
<td>63</td>
</tr>
<tr>
<td>IV</td>
<td>-</td>
<td>11</td>
<td>16</td>
</tr>
<tr>
<td>Not available</td>
<td>3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Smoking history</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never-smokers</td>
<td>20</td>
<td>47</td>
<td>57</td>
</tr>
<tr>
<td>Smokers</td>
<td>104</td>
<td>334</td>
<td>304</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>53</td>
<td>254</td>
<td>173</td>
</tr>
<tr>
<td>Female</td>
<td>71</td>
<td>190</td>
<td>200</td>
</tr>
<tr>
<td><strong>Mutation status</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EGFR-mutated</td>
<td>12</td>
<td>-</td>
<td>49*</td>
</tr>
<tr>
<td>KRAS-mutated</td>
<td>24</td>
<td>-</td>
<td>100*</td>
</tr>
<tr>
<td><strong>Patient outcome</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outcome type**</td>
<td>OS (121/83)</td>
<td>RFS (198/155)</td>
<td>-</td>
</tr>
<tr>
<td><strong>Evaluable for</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CNAs</td>
<td>X</td>
<td>X***</td>
<td>X</td>
</tr>
<tr>
<td>Gene expression</td>
<td>X</td>
<td>-</td>
<td>X</td>
</tr>
<tr>
<td>Mutation spectrum</td>
<td>-</td>
<td>-</td>
<td>X</td>
</tr>
</tbody>
</table>

* non-silent mutations from Mutation Annotation Format (MAF) file.

** OS: overall survival, RFS: relapse-free survival. Nbr patients with outcome data (NSCLC / Adenocarcinoma)

*** Only CN-FGA
Figure legends

Figure 1. DNA methylation patterns in lung cancer. (A) Distribution of average beta values for 473864 CpGs stratified by Illumina gene location across the 124 tumors in the discovery cohort. TSS: transcription start site. (B) Spearman correlation of DNA methylation and gene expression for 9334 gene matching CpGs in 77 adenocarcinomas from the discovery cohort, stratified by Illumina gene location. Top axis indicates number of CpGs per group. (C) Spearman correlation for the 9334 CpGs grouped according to the human embryonic stem cell (H1hESC) chromatin state track (12). Digits correspond to track state id. The number of CpGs per group are indicated on top of the panel. (D) 4136 CpGs were selected based on variation in at least 13 out of 124 tumor cases compared to 12 normal lung tissues and grouped according to Illumina CpG island annotations. Orange bars correspond to CpGs hypermethylated in tumors, blue bars correspond to CpGs hypomethylated in tumors, and black bars correspond to distribution of all CpGs on the Illumina platform. (E) The 4136 CpGs grouped according to the human embryonic stem cell (H1hESC) chromatin state track.

Figure 2. Identification of five DNA methylation subtypes in the discovery cohort. (A) DNA methylation subtypes in 124 lung cancers based on bootstrap clustering of 4136 variant CpGs. Heatmap displays beta values (rows) from unmethylated (blue) to methylated (yellow) for three sample groups (columns): 124 tumors divided into five subtypes by bootstrap clustering, 12 matched normal lung tissues, and blood leukocytes, with associated clinical characteristics and reported adenocarcinoma (AC) and squamous cell carcinoma (SqCC) gene expression...
phenotypes (9, 21). Left hand CpG tracks: CpG island track, black=island, grey=shore/shelf, white=open sea, H1hESC track (12) (embryonic stem cell chromatin state): purple=poised promoter, red=active promoter, yellow=enhancer, green=transcribed, blue=insulator, white=heterochromatin. Sample annotations: black=yes, grey=no. (B) Global promoter hypermethylation (left) and global hypermethylation (right) score for methylation clusters (based on all filtered CpGs on the platform). (C) Box plots of DNA methylation for 629 CpGs matching repetitive elements from the set of 4136 for each tumor in the discovery cohort across epitypes. Tumors are colored according to epitype as in A, with exception for ES5 (gray).

Figure 3. Clinicopathological and gene expression characteristics of adenocarcinoma epitypes. (A) Distribution of adenocarcinomas in epitypes in the discovery, Sandoval, and TCGA cohorts. (B) Distribution of never-smokers with adenocarcinoma in epitypes in the discovery, Sandoval, and TCGA cohorts. P-values calculated using Fisher’s exact test. (C) Distribution of EGFR (left) and KRAS mutations (right) in the epitypes in the discovery and TCGA cohorts. (D) Association of reported gene expression adenocarcinoma (AC) subtypes (bronchioid, magnoid, and squamoid) (9) with epitypes in the discovery (top) and TCGA (bottom) cohorts. Combinations with clear enrichment highlighted in red. P-values were calculated using the Chi-square test. (E) Association of epitypes with gene expression subgroups derived from individual unsupervised consensus clustering (50) of adenocarcinomas in the discovery cohort (top) and TCGA (bottom) cohort. For the discovery cohort, clusters were derived by consensus clustering of genes with log2ratio standard deviation >0.5 across all adenocarcinomas using a four-group cluster solution. The TCGA consensus clusters were derived by unsupervised consensus clustering of the
most varying genes (standard deviation of variation across tumors >1) using a three-
group solution as described in (31). P-values were calculated using the Chi-square
test. (F) Expression of an immune response associated metagene (top) and a
stroma/extracellular matrix (ECM) metagene (bottom) in the discovery cohort for
adenocarcinomas stratified by epitypes. P-values calculated using Kruskal-Wallis test.

Figure 4. Association of adenocarcinoma epitypes with CNAs and mutations. (A)
Total number of mutations for TCGA adenocarcinomas stratified by epitype. (B)
Association of specific mutations with epitypes in TCGA adenocarcinomas.
Mutations were identified by a permutation-based approach applied to a list of
significant mutations in adenocarcinoma identified through MutSigCV analysis as
described ((31) and Supplementary Methods). P-values were calculated using Fisher’s
exact test.

Figure 5. Association of adenocarcinoma epitypes with patient outcome. (A)
Overall survival for NSCLC patients (left) and adenocarcinoma patients (right)
stratified by epitype in the discovery cohort. (B) Relapse-free survival (RFS) for
NSCLC patients (left) and adenocarcinoma patients (right) stratified by epitype in the
Sandoval cohort. In this cohort, no patient included in survival analyses received
adjuvant chemotherapy. P-values calculated using log-rank test.
Figure 2

A) Heatmap showing DNA methylation patterns across different subgroups. The x-axis represents different samples labeled as ES1, ES2, ES3, ES4, and ES5. The y-axis shows different DNA methylation tracks, including H1hESC and Cpg Island track.

B) Box plots showing the distribution of promoter hypermethylation and global hypermethylation scores across different subgroups. The Kruskal-Wallis test is used to compare the groups, with P-values of 3e-12 for promoter hypermethylation and 3e-14 for global hypermethylation.

C) Line plots illustrating DNA methylation beta values across different samples, with hypermethylation indicated by different colors.
Figure 3

(A) Sandoval p=0.06

(B) TCGA p=0.11

(C) Discovery, p=0.01

(D) TCGA P= 2e−28

(E) TCGA P= 4e−13

(F) Research. on July 15, 2017. © 2014 American Association for Cancer Research.
Figure 4

(A) Box plots showing the number of mutations for ES1, ES2, ES4, and ES5. The Kruskal-Wallis test was performed, with a significance level of $P = 7 \times 10^{-11}$.

(B) Table comparing mutation frequencies in ES1, ES2, ES4, and ES5.

<table>
<thead>
<tr>
<th></th>
<th>ES1</th>
<th>ES2</th>
<th>ES4</th>
<th>ES5</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP53 mutation</td>
<td>59%</td>
<td>65%</td>
<td>56%</td>
<td>28%</td>
</tr>
<tr>
<td>STK11 mutation</td>
<td>33%</td>
<td>18%</td>
<td>6%</td>
<td>10%</td>
</tr>
<tr>
<td>KEAP1 mutation</td>
<td>30%</td>
<td>32%</td>
<td>7%</td>
<td>9%</td>
</tr>
<tr>
<td>SMARCA4 mutation</td>
<td>7%</td>
<td>18%</td>
<td>5%</td>
<td>4%</td>
</tr>
</tbody>
</table>

P-values:
- For TP53 mutation: $P = 2 \times 10^{-6}$
- For STK11 mutation: $P = 6 \times 10^{-5}$
- For KEAP1 mutation: $P = 2 \times 10^{-7}$
- For SMARCA4 mutation: $P = 0.002$
**Figure 5**

(A) NSCLC Discovery

- **Overall survival (%)**
  - ES1: n=17
  - ES2: n=27
  - ES3: n=7
  - ES4: n=31
  - ES5: n=39
  - *P = 0.06*

- **RFS (%)**
  - *P = 0.03*

(B) NSCLC Sandoval

- **Overall survival (%)**
  - ES1: n=12
  - ES2: n=52
  - ES4: n=46
  - ES5: n=88
  - *P = 0.01*

- **RFS (%)**
  - *P = 0.01*

(C) Adenocarcinoma Discovery

- **Overall survival (%)**
  - ES1: n=10
  - ES2: n=12
  - ES4: n=30
  - ES5: n=31
  - *P = 0.01*

- **RFS (%)**
  - *P = 0.03*

(D) Adenocarcinoma Sandoval

- **Overall survival (%)**
  - ES1: n=7
  - ES2: n=27
  - ES4: n=45
  - ES5: n=76

- **RFS (%)**
  - *P = 0.03*
Clinical Cancer Research

Genome-wide DNA methylation analysis of lung carcinoma reveals one neuorendocrine and four adenocarcinoma epitypes associated with patient outcome

Anna Karlsson, Mats Jonsson, Martin Lauss, et al.

Clin Cancer Res  Published OnlineFirst October 2, 2014.

Updated version
Access the most recent version of this article at:
doi:10.1158/1078-0432.CCR-14-1087

Supplementary Material
Access the most recent supplemental material at:
http://clincancerres.aacrjournals.org/content/suppl/2014/10/03/1078-0432.CCR-14-1087.DC1

Author Manuscript
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

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.