Lung adenocarcinoma subtypes based on expression of human airway basal cell genes

Tomoya Fukui[1,5], Renat Shaykhiev[1,5], Francisco Agosto-Perez[2], Jason G. Mezey[1,2], Robert J. Downey[3], William D. Travis[4] and Ronald G. Crystal[1]

Affiliations: [1]Dept of Genetic Medicine, Weill Cornell Medical College, New York, NY, [2]Dept of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY, [3]Thoracic Service, Dept of Surgery, Memorial Sloan-Kettering Cancer Center, New York, NY, and [4]Surgical Pathology Diagnostic Services, Thoracic, Dept of Pathology, Memorial Sloan-Kettering Cancer Center, New York, NY, USA. [5]Both investigators contributed equally.

Correspondence: R.G. Crystal, Dept of Genetic Medicine, Weill Cornell Medical College, 1300 York Avenue, Box 96, New York, NY 10065, USA. E-mail: geneticmedicine@med.cornell.edu

ABSTRACT Lung cancer, including lung adenocarcinoma, is a heterogeneous disease, which evolves from molecular alterations in the airway epithelium. This study explores whether a subtype of lung adenocarcinomas expresses the unique molecular features of human airway basal cells (BCs), and how expression of the airway BC features correlates with the molecular, pathological and clinical phenotype of lung adenocarcinoma.

Three independent lung adenocarcinoma data sets were analysed for expression of genes that constitute the airway BC signature. Expression of the BC signature in lung adenocarcinoma was then correlated to clinical and biological parameters.

Remarkable enrichment of airway BC signature genes was found in lung adenocarcinomas. A subset of lung adenocarcinomas (BC-high adenocarcinoma) exhibited high expression of BC signature genes in association with poorer tumour grade, higher frequency of vascular invasion and shorter survival than adenocarcinomas with lower expression of these genes. At the molecular level, BC-high adenocarcinomas displayed a higher frequency of *KRAS* mutations, activation of transcriptional networks and pathways related to cell cycle, extracellular matrix organisation, and a distinct differentiation pattern with suppression of ciliated and exocrine bronchiolar cell (Clara cell)-related genes.

Activation of the airway BC programme is a molecular feature of a distinct, aggressive subtype of lung adenocarcinoma.

@ERSpublications
High expression of airway basal cell genes contributes to the molecular phenotype of aggressive lung adenocarcinomas http://ow.ly/nIdJC

This article has supplementary material available from www.erj.ersjournals.com

Received: Sept 11 2012 | Accepted after revision: Jan 21 2013 | First published online: May 03 2013

Support statement: These studies were supported, in part, by National Institutes of Health grants P50 HL084936, T32 HL094284 and UL1 RR02499, and the Starr Foundation/Starr Cancer Consortium. R. Shaykhiev is supported, in part, by the Parker B. Francis Foundation.

Conflict of interest: None declared.

Introduction

Lung cancer, the leading cause of cancer mortality worldwide, is a heterogeneous disease that evolves from molecular alterations in the airway epithelium mediated by environmental oncogenic stress, primarily cigarette smoking [1–3]. The human airway epithelium is a pseudostratified layer dominated by ciliated cells together with secretory, intermediate, basal cells (BCs), and rare neuroendocrine cells [4].

The specific contribution of these individual cell types of the airway epithelium to lung cancer heterogeneity is not well understood. Airway BCs, the stem/progenitor cell population of the airway epithelium [5], are considered the candidate ''cell of origin'' of lung squamous cell carcinoma, in part because airway BCs are probably the source of squamous cell metaplasia, the squamous cell carcinoma-related potential pre-neoplastic lesion [1]. In contrast, the cellular origin of lung adenocarcinoma is not clear [1]. Centrally located adenocarcinomas are thought to arise from the surface or glandular epithelium of bronchi [6]. By contrast, an exocrine bronchiolar cell (Clara cell)- and type II pneumocyte-associated differentiation pattern, also known as the ''terminal respiratory unit'' (TRU), has been observed in peripheral adenocarcinoma [7, 8]. Subpopulations of exocrine bronchiolar cells, a secretory cell type present throughout the airway epithelium in mice [9], but limited to small airways in humans [10], have been related to lung adenocarcinoma development in murine models [11]. However, the contribution of other cell types, such as BC-like progenitors, to human lung adenocarcinoma has also been proposed [1, 12].

In the present study, using our recent description of the human airway BC transcriptome [13], we analysed the contribution of the unique molecular features of airway BCs to the molecular and clinical phenotype of lung adenocarcinoma. The data provides evidence for a subtype of lung adenocarcinoma that expresses high levels of airway BC genes in association with an aggressive clinical phenotype.

Methods

Additional details of the methods used can be found in the online supplementary material.

Lung adenocarcinoma data sets

Three independent lung adenocarcinoma cohorts were analysed: a primary cohort (182 out of 199 subjects originally described by CHITALE *et al* [14], which was re-evaluated histologically and updated with recent clinical information) and two validation cohorts, one described by BILD *et al.* [15] (58 subjects) and one by SHEDDEN *et al.* [16] (327 out of 442 subjects, *i.e.* excluding 104 subjects analysed in Memorial Sloan Kettering Cancer Center (MSKCC) (New York, NY, USA), the majority of which are present in the primary cohort, and 11 large cell neuroendocrine carcinoma samples identified based on pathologic re-evaluation [17]). Patient characteristics are summarised in online supplementary table I. A diagram representing the experimental flow of the study is shown in supplementary figure S1.

Analysis of airway BC signature expression in human lung adenocarcinoma

The airway BC signature (862 genes) was previously characterised in our laboratory [13]. To analyse gene enrichment, microarray data was normalised by chip and then median expression levels for all genes across all samples was determined. Median levels for each gene were compared to the median level for all 862 airway BC signature genes, the non-BC signature, which had significantly higher expression in the complete large airway epithelium *versus* purified BC based on the genome-wide microarray comparison (criteria for high expression: fold-change $\geqslant 5$, $p<0.01$ with Benjamini–Hochberg correction), and 50 random 862-gene sets (selected from the Affymetrix (Santa Clara, CA, USA) HG-U133A genome using Excel (Microsoft Corporation, Redmond, WA, USA) RAND function).

To compare the expression of the airway BC signature among various carcinoma subtypes [18–24] with airway BC samples, the data sets were analysed by principal component analysis (PCA) using GeneSpring version 7.3.1 (Agilent Technologies, Santa Clara, CA).

A BC index (I_{BC}) was calculated for each individual subject as a cumulative measure of the airway BC signature expression as previously described for the complete airway epithelium [25]. Categorization of subjects was performed based on the I_{BC} using the quartile method: individuals within the bottom quartile were categorised as ''BC low'' and individuals within the top quartile were categorised as ''BC high''.

To determine transcriptome differences between BC-high and BC-low adenocarcinomas, we performed genome-wide comparison (criteria for differentially expressed genes: fold-change >2, $p<0.01$ with Benjamini–Hochberg correction). Enrichment of pathways within differentially expressed genes was analysed using the DAVID Bioinformatics Resources 6.7 analytic tool (http://david.abcc.ncifcrf.gov/). To analyse networks for the BC-high adenocarcinoma upregulated genes, co-expressed genes were identified in the upregulated genes using Weighted Correlation Network Analysis and identified network genes (criteria: Spearman correlation $\rho>0.6$, $p<0.05$) were then linked to the airway BC signature genes upregulated in

BC-high lung adenocarcinoma based on the known physical protein–protein interactions and transcriptional regulation using the GNCPro analytic tool (http://gncpro.sabiosciences.com/gncpro/gncpro.php).

Expression of genes associated with the major cell types of the human airway epithelium (ciliated, mucus-secreting, exocrine bronchiolar and neuroendocrine cells) and epithelial–mesenchymal transition (EMT) were compared in the lung adenocarcinoma subtypes of the primary cohort. To compare the expression of the airway BC signature in lung adenocarcinomas to squamous cell carcinomas, the dataset containing 58 adenocarcinomas and 53 squamous cell carcinomas described by BILD *et al.* [15] was analysed.

Survival analysis

To assess the relationship of expression of the airway BC signature on survival of patients with lung adenocarcinoma, we first identified poor survival-associated genes by genome-wide comparison between adenocarcinoma patients with <2-year overall survival (poor survivors) *versus* those with >5-year overall survival in the primary cohort (criteria for differentially expressed genes: $p<0.05$ with Benjamini–Hochberg correction). All survival analyses were performed using the Kaplan–Meier method. Survival between the adenocarcinoma subtypes was compared using the log-rank test. Multivariate analysis was performed using a Cox proportional hazard model.

Immunohistochemical analysis

Biopsy samples were independently collected from adenocarcinoma patients undergoing lung resection according to the protocol and informed consent approved by the MSKCC Institutional Review Board. Categorisation of the lung adenocarcinoma samples used in immunohistochemistry into BC high and low was made using the index method, as described earlier, based on TaqMan PCR analysis (Applied Biosystems, Foster City, CA, USA) of the expression of the top 10 genes with >85% sensitivity for BC-high adenocarcinoma (in all three independent lung adenocarcinoma data sets; gene list is presented in online supplementary table II). Immunohistochemical analysis was performed to validate differential expression of selected proteins between BC-low and -high adenocarcinoma [13]. The only modification was that the samples were incubated with primary antibodies against tumour protein 63 (TP63) (2 $\mu g \cdot mL^{-1}$; Santa Cruz Biotechnology, Santa Cruz, CA, USA) and thyroid transcription factor-1 (TTF-1) (3 $\mu g \cdot mL^{-1}$; DAKO, Carpinteria, CA, USA) for 2 h at 37°C. Commercially available normal lung and lung squamous cell carcinoma tissue samples (US Biomax Inc., Rockville, MD, USA) were used for comparative analysis.

Statistical analysis

All analyses, except for the microarray data, were performed using the SPSS statistical package (SPSS Inc, Chicago, IL, USA). The relationship between the I_{BC-} and the *NKX2-1* gene expression was analysed in the primary adenocarcinoma cohort using Pearson correlation analysis. The relationship between the groups was assessed using the Chi-squared or Mann–Whitney test. Analysis of the microarray data was performed as specified earlier using GeneSpring version 7.3.1.

Results

Airway BC signature is enriched in lung adenocarcinoma

To provide comprehensive view on the expression of airway BC molecular features in lung adenocarcinoma, expression of the 862-gene airway BC signature (online supplementary gene list I) was analysed. Of the 862 airway BC signature genes, 420 (48.7%) were among the highly expressed lung adenocarcinoma genes (expression level more than twice the median for all genes), whereas only 118 (22%) out of 544 non-BC signature genes (online supplementary gene list II) and <35% of genes in randomly selected gene sets were among highly expressed lung adenocarcinoma genes (fig. 1a). By contrast, <20% of the BC-signature genes contributed to the genes with low expression in lung adenocarcinoma (expression level less than half the median for all genes) as compared to >35% of the non-BC genes and those from the randomly selected data sets. The enrichment of the BC signature genes in lung adenocarcinoma was validated using two independent cohorts (fig. 1b and c). Combined analysis of all three cohorts revealed statistically significant enrichment of the airway BC signature genes among the highly expressed lung adenocarcinoma genes *versus* non-BC genes ($p<0.0006$) and *versus* randomly selected gene sets ($p<0.02$) (online supplementary table III).

Airway BC signature is upregulated in a subset of lung adenocarcinoma

Next, we asked whether the pattern of airway BC signature expression in lung adenocarcinoma is shared by other carcinomas or relatively unique to this type of lung cancer. The PCA revealed that lung squamous cell carcinoma [20] and all three lung adenocarcinoma data sets [20–22] exhibited similar patterns, with clustering closer to airway BC samples than colorectal [18], breast [19], hepatocellular [23] and pancreas

DOI: 10.1183/09031936.00144012

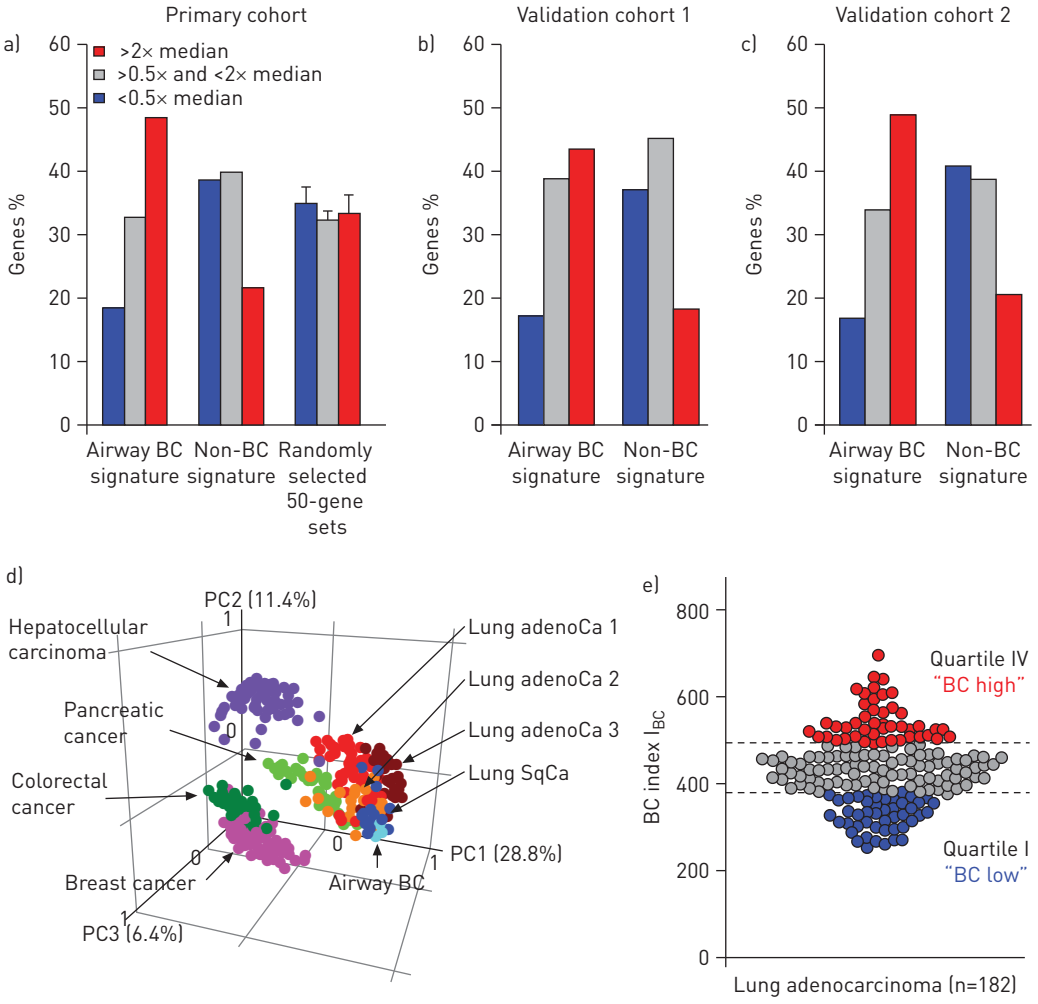

FIGURE 1 Expression of the airway basal cell (BC) signature genes in human lung adenocarcinoma (adenoCa). Frequency of the airway BC signature genes, non-BC signature genes and the genes of the 50 random 862-gene sets contributing to the genes with high expression in lung adenoCa (expression level more than twice the median for all expressed genes; red), genes with low expression in lung adenoCa (expression level less than half the median for all expressed genes; blue) and genes with intermediate expression in lung adenoCa (remaining lung adenoCa-expressed genes; grey) in the a) primary cohort [14], and validation cohorts b) 1 [15] and c) 2 [16]. See online supplementary table III for details. d) Principal component analysis comparing various types of human carcinomas and airway BCs based on expression of the airway BC signature. Analysed data sets include lung adenoCa 1 from DING *et al.* [22] (red; n=68); lung adenoCa 2 from HOU *et al.* [21] (dark red; n=40), adenoCa 3 from KUNER *et al.* [20] (orange; n=40), lung squamous cell carcinoma (SqCa) from KUNER *et al.* [20] (dark blue; n=18), colorectal cancer from SMITH *et al.* [18] (dark green; n=55), breast cancer from LU *et al.* [19] (pink; n=129), hepatocellular carcinoma from CHIANG *et al.* [23] (purple; n=91), pancreatic cancer from BADEA *et al.* [24] (light green; n=39) and airway BC samples from healthy nonsmokers (light blue; n=4). Each circle represents an individual sample. The per cent contributions of the first three principal components (PCs) to the observed variability are indicated. e) Categorisation of lung adenoCa into high and low airway BC gene expressers. The BC index (I_{BC}) is based on a number of airway BC signature genes expressed above the median level in lung adenoCa subjects (n=182) [14]. AdenoCa subjects were divided into BC-high (quartile IV; red), BC-intermediate (quartiles II–III; grey) and BC-low (quartile I; blue) subtypes.

[24] cancers (fig. 1d). The majority of the lung squamous cell carcinoma samples displayed similarity to the airway BC gene expression pattern, whereas the lung adenocarcinoma was more heterogeneous.

To further explore the heterogeneity of lung adenocarcinoma based on the airway BC signature expression, I_{BC} was developed as a cumulative gene expression parameter. Consistent with the PCA data above, the analysis revealed remarkable heterogeneity of lung adenocarcinoma patients based on the airway BC signature expression (fig. 1e). Based on the I_{BC}, BC-high (top quartile) and BC-low (bottom quartile) adenocarcinoma subtypes were identified (fig. 1e).

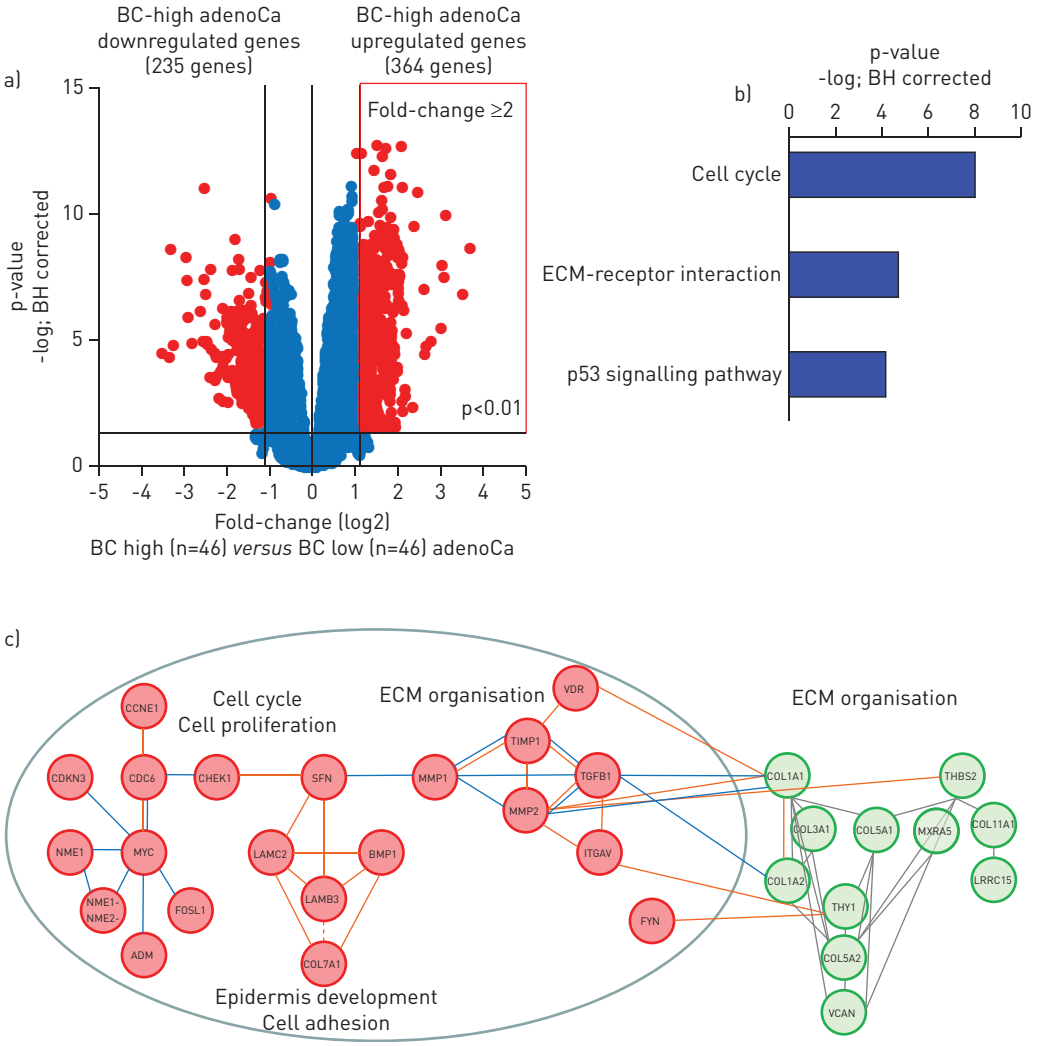

FIGURE 2 Differentially expressed genes between basal cell (BC)-high lung adenocarcinoma (adenoCa) and BC-low adenoCa. a) Volcano plot of genome-wide comparison between BC-high adenoCa (n=46) and BC-low adenoCa (n=46) in the primary lung adenoCa cohort (n=182). Red: significant genes (fold change ⩾2; p<0.01 with Benjamini–Hochberg (BH) correction); blue: genes with no significant difference between the groups. b) Kyoto Encyclopedia of Genes and Genomes pathway analysis of significantly (p<0.05 with BH correction) enriched in BC-high adenoCa upregulated genes. c) Molecular networks enriched within BC-high adenoCa upregulated genes identified using the Weighted Correlation Network Analysis (green: co-expressed BC-high lung adenoCa genes within the primary data set) and connected to the BC signature genes upregulated in BC-high adenoCa (red) using the GCNPro analytic tool (orange: transcriptional regulation; blue: physical protein–protein interactions; see Methods section for details). ECM: extracellular matrix; VDR: vitamin D receptor; CCN: cyclin E1; TIMP1: tissue inhibitor of metalloproteases 1; CDKN3: cyclin-dependent kinase inhibitor 3; CDC6: cell division cycle 6; CHEK1: checkpoint kinase 1; SFN: stratifin; MMP1: matrix metalloprotease 1; TGFB1: transforming growth factor-β1; COL1A1: collagen Iα1; THBS2: thrombospondin 2; MMP2: matrix metalloprotease 2; COL3A1: collagen IIIα1; COL5A1: collagen Vα1; MXRA5: matrix remodelling-associated 5; COL11A1: collagen XIα1; NME1: NME/NM23 nucleoside diphosphate kinase 1; LAMC2: laminin-γ2; BMP1: bone morphogenetic protein 1; ITGAV: integrin-α_V; COL1A2: collagen Iα2; LRRC15: leucine-rich repeat-containing 15; LAMB3: laminin-β3; NME1-NME2: *NME1–NME2* gene readthrough; FOSL1: Fos-like antigen 1; COL5A2: collagen Vα2; ADM: adrenomedullin; COL7A1: collagen VIIα1; VCAN: versican.

BC-high lung adenocarcinoma exhibits distinct biological phenotype

To determine biological pathways and patterns enriched in BC-high adenocarcinoma, we first performed genome-wide comparison of the BC-high *versus* BC-low adenocarcinoma (fig. 2a and online supplementary gene list III). Among the 364 genes up-regulated in BC-high adenocarcinoma, there was significant enrichment of the biological pathways related to the cell cycle, extracellular matrix (ECM)–receptor interaction and the p53 signalling pathway (fig. 2b). Consistent with the pathway analysis, the network analysis of the BC-high adenocarcinoma upregulated genes revealed enrichment of the transcriptional

DOI: 10.1183/09031936.00144012

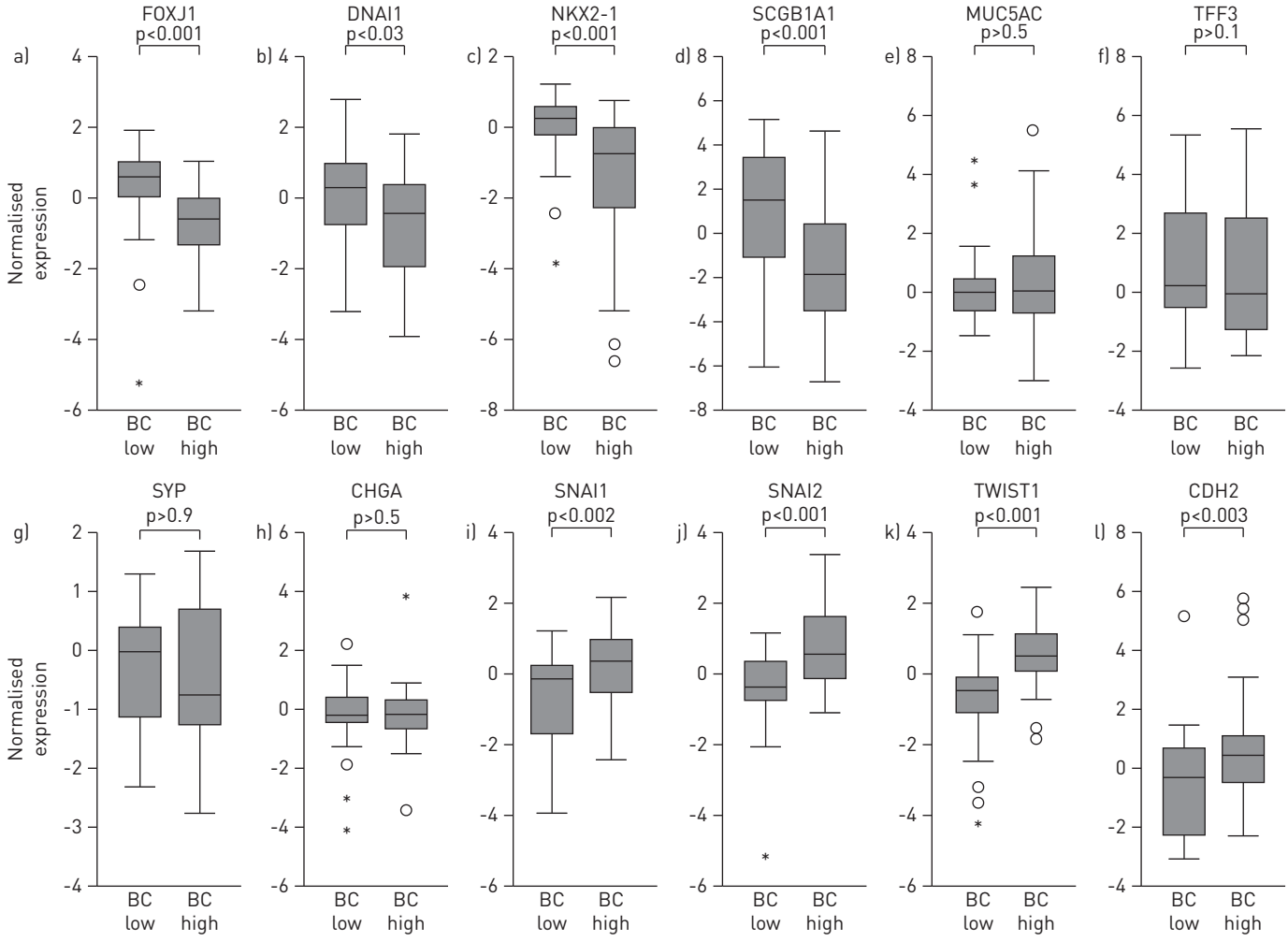

FIGURE 3 Examples of expression of differentiation-associated molecular patterns in basal cell (BC)-high adenocarcinoma compared to BC-low adenocarcinoma. Ciliated cell genes: forkhead box J1 (*FOXJ1*) and dynein axonemal intermediate chain 1 (*DNAI1*); exocrine bronchiolar cell (Clara cell) genes: NK2 homeobox 1 (*NKX2-1*) and secretoglobin 1A1 (*SCGB1A1*); mucin-producing secretory cell-related genes: mucin 5AC (*MUC5AC*) and trefoil factor 3 (*TFF3*); neuroendocrine cell genes: synaptophysin (*SYP*) and chromogranin A (*CHGA*); genes associated with epithelial–mesenchymal transition: snail homolog 1 (*SNAI1*), snail homolog 2 (*SNAI2*), twist homolog 1 (*TWIST1*), and N-cadherin (*CDH2*). In all panels, $\log_2$-transformed normalised gene expression levels based on the microarray analysis are shown; n=46 in each group. Outliers are indicated on the basis of interquartile range (IQR): circles, -1.5 × IQR to 3 × IQR; stars: >3 or <3 × IQR.

network elements related to the ECM organisation (fig. 2c). The BC-high adenocarcinoma-enriched co-expressed ECM network components were interaction partners of the BC signature genes regulating epithelial–mesenchymal interactions and lung tissue homeostasis, including transforming growth factor-β1 (TGFB1), metalloprotease (MMP)1 and MMP2, tissue inhibitor of metalloproteases (TIMP)1, integrin-α_V (ITGAV) and vitamin D receptor (VDR) (fig. 2c).

BC-high adenocarcinoma displayed significant downregulation of genes associated with differentiation of the major cell types of the small airway epithelium, including ciliated cells (forkhead box J1 (*FOXJ1*) and dynein axonemal intermediate chain 1 (*DNAI1*)) and exocrine bronchiolar cells (NK2 homeobox 1 (*NKX2-1*) and secretoglobin 1A1 (*SCGB1A*)). Expression of genes typical for mucus-secreting cells and neuroendocrine cells was not different between these two subtypes. There was a negative correlation between the I_{BC} and *NKX2-1* gene expression (online supplementary fig. S2). Consistent with this observation, there was a trend for a lower expression of the *NKX2-1*-encoded TTF-1 in BC-high adenocarcinoma, although it was detectable in both adenocarcinoma subtypes, unlike the TTF-1-negative squamous cell carcinomas (online supplementary fig. S3). By contrast to differentiation genes, the expression of genes related to EMT, such as *SNAI1* (snail homologue 1), *SNAI2* (snail homologue 2), *TWIST1* (twist basic helix–loop–helix transcription factor 1) and *CDH2* (N-cadherin) [26], was up-regulated in BC-high, compared with BC-low, adenocarcinoma (fig. 3). Genes related to other lung cancer subtypes, including small cell lung carcinoma, such as those encoding p53, retinoblastoma-1 and L-MYC,

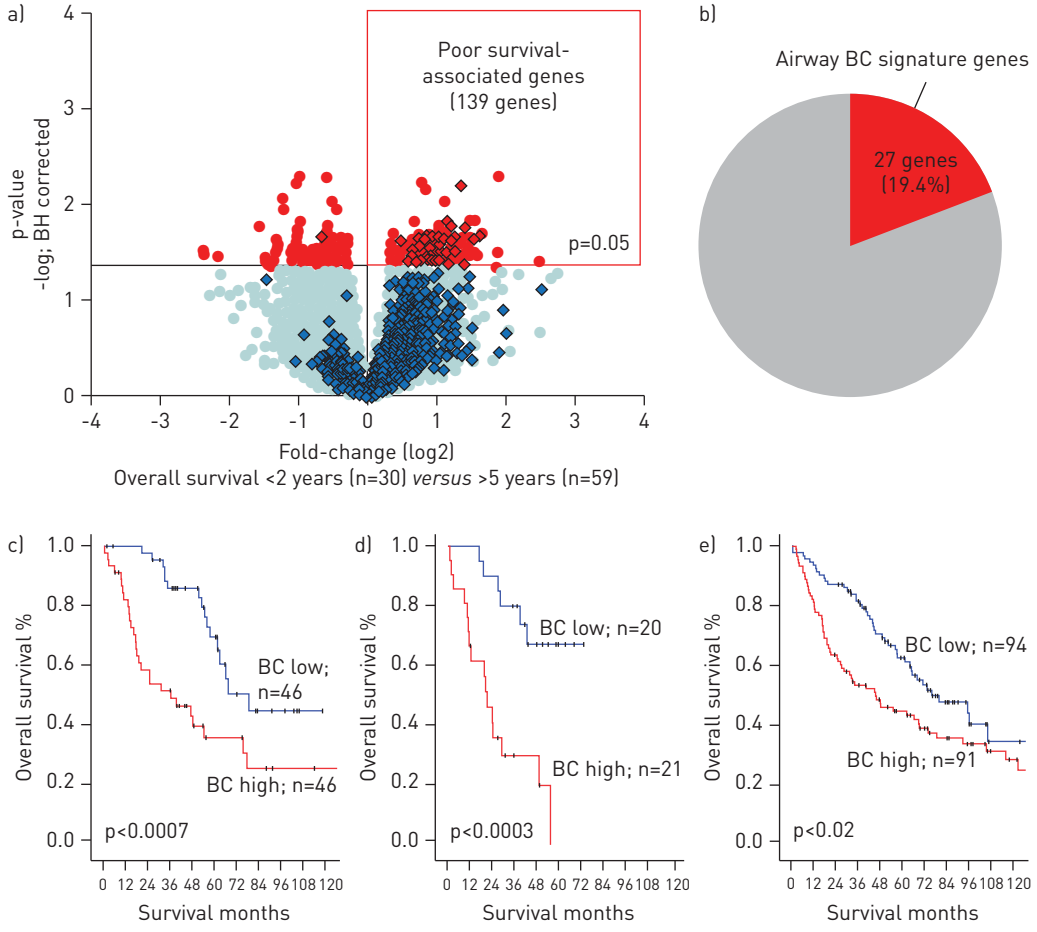

FIGURE 4 Relationship between airway basal cell (BC) signature expression and lung adenocarcinoma patient survival. a) Genome-wide comparison between adenocarcinoma patients with <2-year overall survival (poor survivors; n=30) and those with >5-year overall survival (n=59) in primary lung adenocarcinoma cohort (n=182). Red: significant genes ($p<0.05$ with Benjamini–Hochberg (BH) correction); blue: genes with no significant difference between the groups; diamonds: 862 airway BC signature genes. b) Overlap between the 139 poor survival-associated genes (identified as described above) and airway BC signature genes. c–e) Kaplan–Meier analysis-based estimates of overall survival of BC-high adenocarcinoma patients (red) *versus* BC-low adenocarcinoma patients (blue) from c) the primary cohort of CHITALE *et al.* [14], and the validation cohorts of d) BILD *et al.* [15] and e) SHEDDEN *et al.* [16]. P-values were determined by the log-rank test; the number of individuals in each group is indicated.

were not differentially expressed between the BC-high and BC-low adenocarcinoma subtypes (online supplementary fig. S4).

BC-high adenocarcinoma exhibits an aggressive clinical phenotype

Upregulation of cell cycle-related and EMT genes and suppression of the differentiation-related gene expression programme in BC-high adenocarcinoma suggest that high expression of the BC signature in lung adenocarcinoma may be associated with more aggressive tumour phenotype. The analysis revealed that among 139 genes associated with poor survival in lung adenocarcinoma (fig. 4a), ~20% genes are BC signature genes (fig. 4b and online supplementary gene list IV).

Compared to BC-low adenocarcinoma, BC-high adenocarcinoma was characterised by poorer differentiation ($p<0.001$), lower frequency of prognostically favourable adenocarcinoma with a lepidic pattern (formerly bronchoalveolar carcinoma; $p<0.001$) and higher frequency of vascular invasion ($p<0.004$). At the molecular level, BC-high adenocarcinoma exhibited a significantly higher frequency of *KRAS* mutations and lower frequency of *EGFR* (epidermal growth factor receptor) mutations. Consistent with *KRAS* mutation status, which are known to be more characteristic of smoking-associated adenocarcinoma [1, 3], there were significantly more smokers among BC-high, compared with BC-low, adenocarcinoma patients (table 1).

DOI: 10.1183/09031936.00144012

TABLE 1 Clinicopathological characterisation of the basal cell (BC)-high and BC-low adenocarcinoma

	BC low	BC high	p-value[#]
Subjects	46	46	
Age years	65.9±12.2	68.5±8.9	>0.2
Sex			>0.8
Male	24	24	
Female	22	22	
Smoking history			<0.04
Never[¶]	14	6	
Ever	31	40	
COPD comorbidity			>0.1
No	40	34	
Yes	6	12	
Pathological stage[+]			<0.05
I	36	27	
II/III	5/5	5/14	
Tumour size cm	3.0±1.4	4.1±2.9	<0.05
Node metastasis			<0.04
No (N0)	37	28	
Yes (N1/N2)	6/3	8/10	
Pathological grade			<0.001
Well/Moderate	19/19	1/21	
Poor	5	21	
Lepidic pattern[§]			<0.001
No	17	39	
Yes	2/27	2/5	
Vascular invasion			<0.004
No	35	18	
Yes	8	18	
***NKX2-1* (TTF-1) expression**[f]	1.20±0.55	0.63±0.48	<0.001
***EGFR*[##]**			<0.03
Wild-type	32	41	
Mutant	14	5	
TP53			>0.2
Wild	38	33	
Mutant	8	13	
***KRAS*[¶¶]**			<0.04
Wild-type	39	30	
Mutant	7	16	

Data are presented as n or mean±SD, unless otherwise stated. COPD: chronic obstructive pulmonary disease; NKX2-1: NK2 homeobox 1; TTF-1: thyroid transcription factor 1; EGFR: epidermal growth factor receptor; TP53: tumour protein 53. [#]: Pearson's Chi-squared test (for categorical variables) and Mann–Whitney test (for continuous values); [¶]: subjects who had never had a smoking habit; [+]: based on [27]; [§]: formerly described as bronchioloalveolar carcinoma; [f]: normalised expression based on the microarray gene expression analysis as described in the Methods section; [##]: patients with lung adenocarcinoma had *EGFR* mutations such as deletion in exon 19 (n=24) and a point mutation (L858R) in exon 21 (n=14); [¶¶]: included G12C (n=17), G12V (n=16), G12D (n=6), G12A (n=4) and G13D (n=1).

Consistent with the remarkable contribution of the BC signature to the poor survival-associated gene set (fig. 4b), individuals with BC-high adenocarcinoma had shorter overall survival compared to those with BC-low adenocarcinoma (median survival 36 *versus* 79 months; log rank p<0.0007) (fig. 4c). Multivariate survival analysis demonstrated that high expression of the airway BC signature was an independent prognostic factor associated with shorter survival (hazard ratio 1.59, 95% CI 1.14–2.22; p<0.008) (table 2).

The prognostic relevance of the airway BC signature expression was validated using two independent lung adenocarcinoma cohorts [15, 16]. The proportion of BC-high adenocarcinoma cases ranged between 36% and 28%, compared to 25% in the primary cohort. Similar to the primary cohort, the BC-high adenocarcinoma cases had significantly shorter overall survival compared to BC-low adenocarcinoma (fig. 4d and e). BC-high adenocarcinoma was also associated with shorter disease-free survival as compared to BC-low adenocarcinoma (supplementary fig. S5).

TABLE 2 Multivariate Cox regression analyses including the category associated with the airway basal cell (BC) signature

	HR for overall survival (95% CI)	p-value
Age	1.07 (1.03–1.11)	<0.002
Sex	1.14 (0.58–2.22)	>0.7
Smoking status	1.27 (0.46–3.51)	>0.6
Pathological stage	3.81 (2.01–7.22)	<0.001
Lepidic pattern	1.23 (0.25–6.09)	>0.8
Adjuvant therapy	0.87 (0.36–2.15)	>0.7
Airway BC signature	1.59 (1.14–2.22)	<0.008

In the multivariate analyses, age (continuous variable), sex (male *versus* female), smoking status (never- *versus* ever-smoker), pathological stage (I *versus* II–III) [27], pathological feature (adenocarcinoma with lepidic pattern (formerly bronchioloalveolar carcinoma) *versus* other adenocarcinoma), adjuvant therapy (no *versus* yes) and airway BC signature (low *versus* high) were included as factors. Adjuvant chemotherapy referred to systemic chemotherapy performed pre- and/or post-surgery, including adjuvant chemotherapy (n=20), adjuvant chemoradiotherapy (n=3), induction chemotherapy (n=4), and induction chemotherapy and adjuvant chemoradiotherapy (n=1). HR: hazard ratio.

BC-high lung adenocarcinoma is distinct from lung squamous cell carcinoma

Abnormal activation of some airway BC genes has been previously linked to lung squamous cell carcinoma [12]. Consistent with these prior observations, the overall expression of the airway BC signature was significantly higher in squamous cell carcinoma compared to the adenocarcinoma cohort (fig. 5a). However, there was no significant difference in the overall survival between BC-high and BC-low squamous cell carcinoma individuals (online supplementary fig. S6a and b). Notably, despite that the overall expression of the BC genes was higher in squamous cell carcinoma compared to adenocarcinoma, the overall survival of the squamous cell carcinoma patients was longer compared to the BC-high adenocarcinoma in the analysed cohort (online supplementary fig. S6c).

Next, we asked whether BC-high adenocarcinoma shares airway BC-related molecular features of squamous cell carcinoma. Comparison of expression of airway BC signature genes in BC-high adenocarcinoma with squamous cell carcinoma identified 13% of the airway BC genes with higher expression in BC-high adenocarcinoma and 11% in squamous carcinoma (fig. 5b and online supplementary gene list V). This indicates that BC-high adenocarcinoma is characterised by a distinct pattern of airway BC genes that distinguishes this subtype of lung cancer from squamous cell carcinoma. Among the airway BC genes predominantly up-regulated in BC-high adenocarcinoma were keratin-7 (*KRT7*), the EGFR ligand amphiregulin (*AREG*), ErbB receptor feedback inhibitor 1 (*ERRFI1*) and tissue factor pathway inhibitor 2 (*TFFPI2*) (fig. 5c). By contrast, the classical BC markers keratin-5 (*KRT5*), *TP63*, keratin-5B (*KRT6B*) and keratin-17 (*KRT17*) had significantly higher expression in squamous cell carcinoma compared to BC-high adenocarcinoma (fig. 5c). Consistent with this observation, immunohistochemical analysis revealed that TP63 protein, normally expressed in the airway BC population, was overexpressed in squamous cell carcinoma but not in either adenocarcinoma subtype (online supplementary fig. S3).

Discussion

There is a growing body of evidence that biological heterogeneity of human malignancies is determined by specific populations of tissue-resident cells of origin that, under the influence of a distinct set of oncogenic alterations, contribute to particular clinically relevant phenotypes within each individual histological type of cancer [28]. Based on this concept, we have assessed the molecular and clinical heterogeneity of human lung adenocarcinoma using our recent characterisation of the transcriptome of human airway BCs, the stem/progenitor cell population of the airway epithelium [5, 13]. This analysis led us to the identification of a novel biological subtype of lung adenocarcinoma, designated BC-high adenocarcinoma, characterised by upregulation of a distinct set of airway BC signature genes in association with clinical and pathological features of tumour aggressiveness.

Lung cancer originates from molecular alterations in the airway epithelium [1, 3, 8], a cell population comprised of ciliated, intermediate and secretory cells (goblet cells in the large airways and exocrine bronchiolar cells in the small airways), BCs, and rare neuroendocrine cells [4]. Depending on the unique morphological features of individual subtypes of lung cancer, candidate cell types for the origin of each

DOI: 10.1183/09031936.00144012

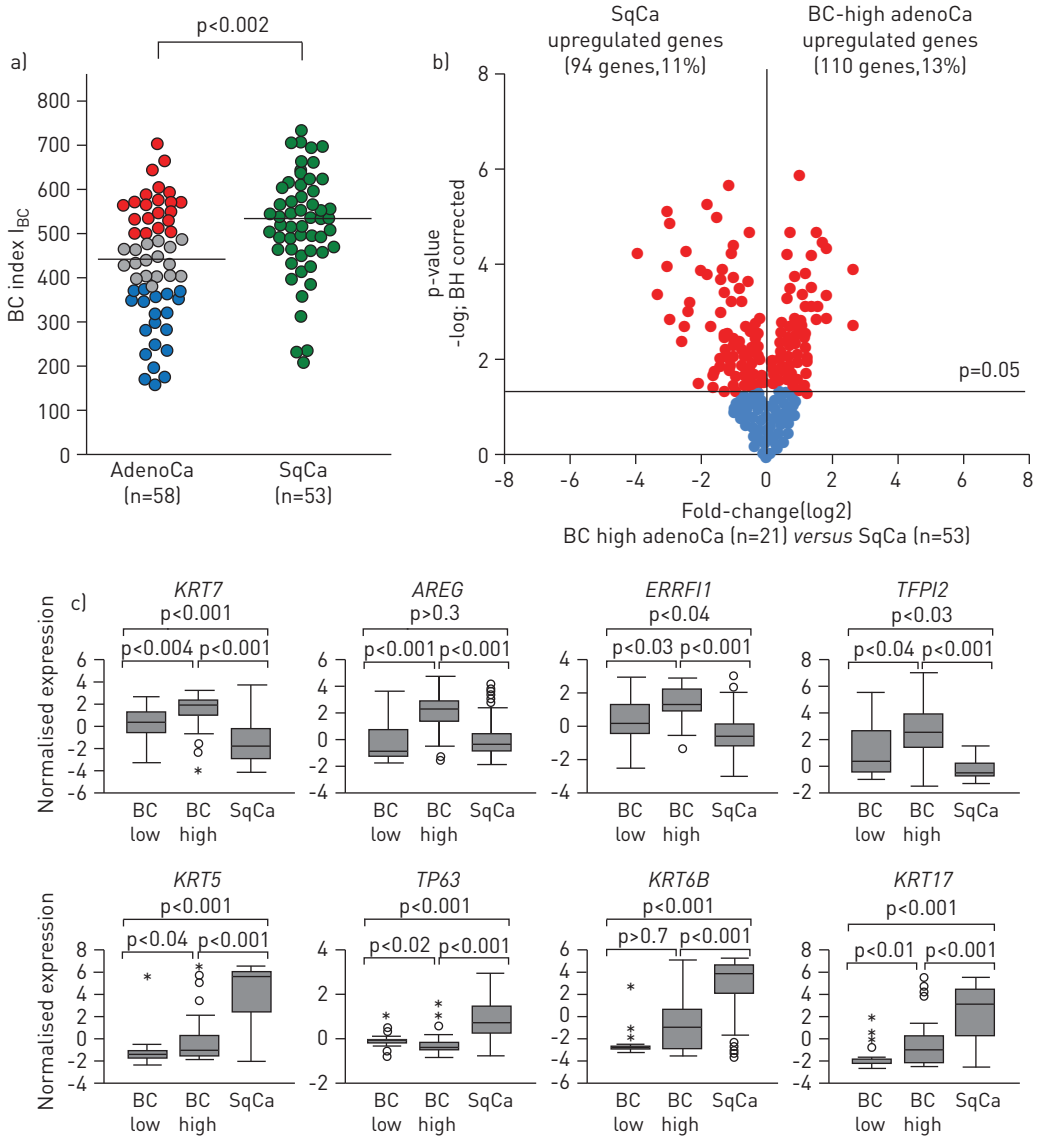

FIGURE 5 Comparative analysis of the airway basal cell (BC) signature expression in lung adenocarcinoma (adenoCa) and lung squamous cell carcinoma (SqCa). a) Lung adenoCa (n=58) and SqCa (n=53) cases from the BILD *et al.* [15] data set were analysed. The BC index (I_{BC}) was calculated based on median levels of adenoCa subjects for each gene. Red: BC-high adenoCa; grey: BC-intermediate adenoCa; blue: BC-low adenoCa; green: SqCa. The median I_{BC} for both types of cancer (436 for adenoCa and 529 for SqCa) are highlighted with horizontal lines; p-value indicated was determined by Mann–Whitney test. b) Volcano plot comparing expression of the airway BC signature genes [13] in BC-high lung adenoCa (n=21) to SqCa (n=53) from the dataset of BILD *et al.* [15]. Red: significant genes ($p<0.05$ with Benjamini–Hochberg (BH) correction); blue: genes with no significant difference between the groups. c) Airway BC signature gene examples differentially expressed in BC-high adenoCa (n=21) *versus* SqCa (n=53) from the data set of BILD *et al.* [15]. In all panels, $\log_2$-transformed normalised gene expression levels are based on the microarray analysis. Outliers were indicated on the basis of interquartile range (IQR): circles: -1.5 × IQR to 3 × IQR; stars: >3 or <3 × IQR. *KRT7*: keratin-7; *AREG*: amphiregulin; *ERRFI1*: ErbB receptor feedback inhibitor 1; *TFPI2*: tissue factor pathway inhibitor 2; *KRT5*: keratin-5; *TP63*: tumour protein 63; *KRT6B*: keratin-6B; *KRT17*: keratin-17.

histological subtype have been proposed. Small cell lung carcinoma and large cell neuroendocrine carcinoma are thought to originate from the neuroendocrine cells [1]. Airway BCs have been considered putative cells of origin of lung squamous cell carcinoma based on the knowledge that airway BCs serve as a source of squamous metaplasia, a histological lesion associated with the early steps of the development of lung squamous cell carcinoma [1], as well as overexpression of selected BC markers such as KRT5 and TP63 in lung squamous cell carcinomas [29]. For lung adenocarcinoma, the cell of origin is not known, although exocrine bronchiolar cells and type II pneumocytes have been proposed as cellular origins of peripheral adenocarcinoma, and cells of the surface and glandular bronchial epithelium as the source of more proximal adenocarcinoma [6]. In the mouse, "bronchioalveolar stem cells", a unique stem cell population at the

bronchioalveolar duct junction sharing features of exocrine bronchiolar and alveolar type II cells, have been implicated in initiation and propagation of lung adenocarcinoma [11]. However, the cellular composition of the human airway epithelium is different from that in mice. In humans, BCs are present throughout the airways, whereas they are virtually absent in the small airways of mice [30], impeding investigation of the role of airway BCs in lung adenocarcinoma development using mouse models.

In the present study, we assessed the biological heterogeneity of lung adenocarcinoma at the transcriptional level by hypothesising that a subtype of lung adenocarcinoma may be derived from airway BCs. Based on the expression of the airway BC signature genes, the data demonstrate that lung adenocarcinoma can be categorised into BC-high and BC-low subtypes, which exhibit remarkably different biological, pathological and clinical characteristics. The data provide insights into the biology of lung adenocarcinoma by demonstrating that the phenotypic diversity of human lung adenocarcinoma can be explained, at least in part, by persistent activation to a greater or lesser degree of the gene expression programme associated with airway BCs.

The molecular patterns associated with BC-high *versus* BC-low adenocarcinoma also provide insights into the mechanisms that could lead to activation of the airway BC programme in a subset of lung adenocarcinoma. First, there is a higher frequency of *KRAS* mutations in BC-high adenocarcinoma. The association of *KRAS* mutations with stem/progenitor cells has been reported with regard to lung cancer development. Endoderm progenitors with *KRAS* mutations exhibit increased proliferation, and fail to differentiate and maintain stem cell characteristics *in vitro* [31]. These data support the concept that the BC-high adenocarcinoma is potentially derived from airway BCs carrying *KRAS* mutations in association with smoking [1, 3]. By contrast, BC-low adenocarcinoma was characterised by a higher frequency of *EGFR* mutations. This is consistent with previous observations that *EGFR* mutations are more frequent in nonsmoking adenocarcinoma patients [3] and are associated with the differentiation pattern known as TRU, with high expression of genes typical of exocrine bronchiolar and alveolar type II cells, such as those encoding TTF-1 and surfactant proteins [7].

Second, BC-high lung adenocarcinoma was enriched in transcriptional pathways and networks related to ECM organisation interacting with various BC signature genes encoding important regulators of homeostatic processes in the lung tissue, including *TGFB1*, *MMP1*, *MMP2*, *TIMP2* (tissue inhibitor of metalloproteases 2), *ITGAV* and *VDR*, as well as the networks associated with epidermis development, cell adhesion, cell cycle and proliferation. Given the anatomical location of BCs immediately above the basement membrane, it is possible that BCs might contribute to the pathogenesis of lung adenocarcinoma by regulation of ECM homeostasis and epithelial–mesenchymal interactions. Activation of distinct growth factor signalling mechanisms, including those related to TGFB1 and/or mediated *via* activation of matrix metalloprotease and integrin signalling at the epithelial–mesenchymal interface, may be responsible for enrichment of the cell adhesion- and cell cycle-related networks, and contribute to more aggressive tumour characteristics observed in BC-high lung adenocarcinoma.

Third, comparative analysis of the differentiation pattern of BC-high *versus* BC-low lung adenocarcinoma revealed that downregulation of genes related to ciliated and exocrine bronchiolar cell differentiation in BC-high adenocarcinoma is accompanied by activation of the EMT transcriptional programme, including induction of transcription factors, such as SNAI1/SNAIL, SNAI2/SLUG, and TWIST1 and CDH2 [26]. Activation of the EMT programme is believed to be an important process promoting cancer invasion and metastasis [32]. Consistent with this concept, BC-high adenocarcinoma exhibited a higher frequency of vascular invasion and lymph node metastasis. Furthermore, EMT has been reported to generate cells with tumorigenic characteristics [33]. The present study reinforces the relationship between EMT and tissue stem cells in the context of lung adenocarcinoma development.

Finally, by comparison to the squamous cell carcinoma, in which BC genes are also highly expressed, we identified that BC-high lung adenocarcinoma exhibits upregulation of a distinct set of the BC signature genes, including the genes related to the EGFR pathway, such as *AREG* and *ERRFI1*. The EGFR pathway is enriched in airway BCs [13] and smoking is known to activate EGFR signalling in the lung epithelial cells without EGFR mutation, inducing cellular processes relevant to lung cancer development [34].

Although the genes contributing to BC-high adenocarcinoma-enriched molecular pathways are not known as classic cancer-driving oncogenes, understanding their interaction is important for the development of novel therapeutic strategies aimed at regulation of tumour cell survival and growth, for example, using the synthetic lethality approach. Such strategies may be particularly beneficial for BC-high lung adenocarcinoma, which is associated with a high frequency of *KRAS* mutations, for which no effective specific targeted approaches have been developed [3]. Targeting interaction between some of the BC-high adenocarcinoma-enriched genes, such as the oncogene *MYC* and cyclin-dependent kinases, has been

DOI: 10.1183/09031936.00144012

recently shown to induce therapeutically relevant synthetic lethality in aggressive BC-like breast cancer [35]. In addition, based on the knowledge that survival, growth and differentiation of BCs is dependent on their adhesion to the ECM components [36] and ECM-related genes are enriched in BC-high adenocarcinomas, it is possible that targeting ECM genes may represent an additional approach to induce synthetic lethality interactions in BC-high adenocarcinoma.

Together, the present study identifies a novel, BC-high subtype of human lung adenocarcinoma, associated with activation of a distinct set of airway BC signature genes and provides transcriptome-based evidence supporting the concept that this aggressive subset of human lung adenocarcinoma is likely derived from the airway BC population.

Acknowledgements

We thank M. Ladanyi (Memorial Sloan-Kettering Cancer Center, New York, NY, USA) for providing the MSKCC adenocarcinoma samples, J. Salit (Weill Cornell Medical College, New York, NY, USA) for supporting microarray analysis and N. Mohamed (Weill Cornell Medical College) for help in preparing this manuscript.

References

1 Wistuba II, Gazdar AF. Lung cancer preneoplasia. *Annu Rev Pathol* 2006; 1: 331–348.
2 Jemal A, Siegel R, Xu J, *et al.* Cancer Statistics, 2010. *CA Cancer J Clin* 2010; 60: 277–300.
3 Herbst RS, Heymach JV, Lippman SM. Lung cancer. *N Engl J Med* 2008; 359: 1367–1380.
4 Crystal RG, Randell SH, Engelhardt JF, *et al.* Airway epithelial cells: current concepts and challenges. *Proc Am Thorac Soc* 2008; 5: 772–777.
5 Rock JR, Onaitis MW, Rawlins EL, *et al.* Basal cells as stem cells of the mouse trachea and human airway epithelium. *Proc Natl Acad Sci USA* 2009; 106: 12771–12775.
6 Travis WD, Brambilla E, Muller-Hermelink HK, *et al.* Pathology and genetics: tumors of the lung, pleura, thymus and heart. Lyon, IARC, 2004.
7 Yatabe Y. *EGFR* mutations and the terminal respiratory unit. *Cancer Metastasis Rev* 2010; 29: 23–36.
8 Travis WD, Brambilla E, Noguchi M, *et al.* International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society international multidisciplinary classification of lung adenocarcinoma. *J Thorac Oncol* 2011; 6: 244–285.
9 Pack RJ, Al-Ugaily LH, Morris G. The cells of the tracheobronchial epithelium of the mouse: a quantitative light and electron microscope study. *J Anat* 1981; 132: 71–84.
10 Boers JE, Ambergen AW, Thunnissen FB. Number and proliferation of clara cells in normal human airway epithelium. *Am J Respir Crit Care Med* 1999; 159: 1585–1591.
11 Kim CF, Jackson EL, Woolfenden AE, *et al.* Identification of bronchioalveolar stem cells in normal lung and lung cancer. *Cell* 2005; 121: 823–835.
12 Ooi AT, Mah V, Nickerson DW, *et al.* Presence of a putative tumor-initiating progenitor cell population predicts poor prognosis in smokers with non-small cell lung cancer. *Cancer Res* 2010; 70: 6639–6648.
13 Hackett NR, Shaykhiev R, Walters MS, *et al.* The human airway epithelial basal cell transcriptome. *PLoS One* 2011; 6: e18378.
14 Chitale D, Gong Y, Taylor BS, *et al.* An integrated genomic analysis of lung cancer reveals loss of DUSP4 in *EGFR*-mutant tumors. *Oncogene* 2009; 28: 2773–2783.
15 Bild AH, Yao G, Chang JT, *et al.* Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 2006; 439: 353–357.
16 Shedden K, Taylor JM, Enkemann SA, *et al.* Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat Med* 2008; 14: 822–827.
17 Bryant CM, Albertus DL, Kim S, *et al.* Clinically relevant characterization of lung adenocarcinoma subtypes based on cellular pathways: an international validation study. *PLoS One* 2010; 5: e11712.
18 Smith JJ, Deane NG, Wu F, *et al.* Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. *Gastroenterology* 2010; 138: 958–968.
19 Lu X, Lu X, Wang ZC, *et al.* Predicting features of breast cancer with gene expression patterns. *Breast Cancer Res Treat* 2008; 108: 191–201.
20 Kuner R, Muley T, Meister M, *et al.* Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes. *Lung Cancer* 2009; 63: 32–38.
21 Hou J, Aerts J, den HB, *et al.* Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PLoS One* 2010; 5: e10312.
22 Ding L, Getz G, Wheeler DA, *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* 2008; 455: 1069–1075.
23 Chiang DY, Villanueva A, Hoshida Y, *et al.* Focal gains of VEGFA and molecular classification of hepatocellular carcinoma. *Cancer Res* 2008; 68: 6779–6788.
24 Badea L, Herlea V, Dima SO, *et al.* Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia. *Hepatogastroenterology* 2008; 55: 2016–2027.
25 Tilley AE, O'Connor TP, Hackett NR, *et al.* Biologic phenotyping of the human small airway epithelial response to cigarette smoking. *PLoS One* 2011; 6: e22798.
26 Kalluri R, Weinberg RA. The basics of epithelial–mesenchymal transition. *J Clin Invest* 2009; 119: 1420–1428.
27 UICC. TNM Classification of Malignant Tumors. 6th Edn. New York, Wiley-Liss Inc., 2002.
28 Visvader JE. Cells of origin in cancer. *Nature* 2011; 469: 314–322.
29 Camilo R, Capelozzi VL, Siqueira SA, *et al.* Expression of P63, keratin 5/6, keratin 7, and surfactant-A in non-small cell lung carcinomas. *Hum Pathol* 2006; 37: 542–546.

30 Rock JR, Randell SH, Hogan BL. Airway basal stem cells: a perspective on their roles in epithelial homeostasis and remodeling. *Dis Model Mech* 2010; 3: 545–556.
31 Quinlan MP, Quatela SE, Philips MR, *et al.* Activated Kras, but not Hras or Nras, may initiate tumors of endodermal origin *via* stem cell expansion. *Mol Cell Biol* 2008; 28: 2659–2674.
32 Thiery JP, Acloque H, Huang RY, *et al.* Epithelial–mesenchymal transitions in development and disease. *Cell* 2009; 139: 871–890.
33 Mani SA, Guo W, Liao MJ, *et al.* The epithelial–mesenchymal transition generates cells with properties of stem cells. *Cell* 2008; 133: 704–715.
34 Filosto S, Becker CR, Goldkorn T. Cigarette smoke induces aberrant EGF receptor activation that mediates lung cancer development and resistance to tyrosine kinase inhibitors. *Mol Cancer Ther* 2012; 11: 795–804.
35 Horiuchi D, Kusdra L, Huskey NE, *et al.* MYC pathway activation in triple-negative breast cancer is synthetic lethal with CDK inhibition. *J Exp Med* 2012; 209: 679–696.
36 Coraux C, Roux J, Jolly T, *et al.* Epithelial cell–extracellular matrix interactions and stem cells in airway epithelial regeneration. *Proc Am Thorac Soc* 2008; 5: 689–694.

DOI: 10.1183/09031936.00144012