Supplemental Methods

Microarray Data Sets

The airway BC signature was previously characterized in our laboratory by genome-wide comparison of the BC transcriptome *vs* transcriptome of the intact airway epithelium derived from healthy nonsmokers, using Affymetrix HG-U133 Plus 2.0 (Affymetrix, Santa Clara, CA) [1]. The 199 lung adenoCa data set originally described by Chitale et al (http://cbio.mskcc.org/Public/lung_array_data/) [2] was used as the primary adenoCa cohort, which was re-reviewed histologically and updated with regard to clinical information (survival data). Of the 199 individuals, there were 182 with available microarray data and detailed clinical information. Those excluded individuals with unspecified gender (n=6), pathological stage IV (n=3) and unknown (n=2) or tumor pathology other than adenoCa (n=6). The transcriptome profiling of this data set was with the Affymetrix HG-U133A (n=87) and HG-U133A 2.0 (n=95) microarrays. Of the 1161 airway BC signature genes [1], 862 were on the microarrays; for the purpose of the present study, these 862 genes are referred to as the "airway BC signature." The 544 non-BC signature genes (i.e., genes up-regulated in the intact airway epithelium *vs.* airway BC) on the microarray were also analyzed.

To visualize similarity and differences between various carcinoma subtypes, lung adenoCa, lung SqCa and airway BC were compared using principal component analysis (PCA; GeneSpring software) based on expression of the airway BC signature [1]. All cancer data sets used for this analysis were based on the Affymetrix HG-U133 Plus 2.0 array and are publically available at the Gene Expression Omnibus (GEO), including: 68 lung adenoCa (GSE12667) [3], 40 lung adenoCa (GSE19188) [4], 40 lung adenoCa and 18 lung SqCa (GSE10245) [5], 55 colorectal cancer (GSE17537) [6], 129 breast cancer (GSE5460) [7], 91 hepatocellular carcinoma (GSE9843) [8] and 39 pancreatic cancer (GSE15471) [9].

To validate the relationship of the expression of airway BC signature genes to survival for adenoCa, two independent lung adenoCa data sets were assessed, including 58 adenoCa (validation cohort 1, http://data.cgt.duke.edu/oncogene.php and GSE3141) [10] and 327 of 442 adenoCa (validation cohort 2, https://caarraydb.nci.nih.gov/caarray/publicExperimentDetailAction.do?ex) [11], excluding 104 subjects analyzed in Memorial Sloan Kettering Cancer Center, the majority of which are present in the primary cohort, and 11 large cell neuroendocrine carcinoma samples identified based on pathologic re-evaluation [12].

Calculation of the BC index (I_{BC}) for each individual subject

For each airway BC signature gene, the median expression level was determined with respect to each array. Then the I_{BC} was calculated for each subject as a number of the airway BC genes having expression levels higher than median level in the analyzed cohort using the formula:

$$I_{BC} = \sum_{n=1}^{862} E_n$$

where E1 had a value of 1 if the expression level for gene 1 was >median level of adenoCa subjects or had a value of 0 if the expression level is ≤median level of adenoCa subjects; E2 is the index for gene 2, etc.

Comparison of Differentiation-associated Molecular Patterns in BC-high *vs* BC-low AdenoCa

To determine differences in the expression of differentiation-associated molecular features in BC-high *vs* BC-low adenoCa, expression of genes associated with the major cell types of the human airway epithelium were compared in the lung adenoCa subtypes of the primary cohort. The genes assessed included those associated with ciliated cells [forkhead box J1

(FOXJ1) and dynein axonemal intermediate chain 1 (DNAI1)]; mucus-secreting cells [mucin 5AC (MUC5AC) and trefoil factor 3 (TFF3)] [13]; Clara cells [NK2 homeobox 1 (NKX2-1) and secretoglobin 1A1 (SCGB1A)] [14], and neuroendocrine cells [synaptophysin (SYP) and chromogranin A (CHGA)] [15]. In addition, expression of genes related to epithelial-mesenchymal transition (EMT), including snail homolog 1(SNAI1), snail homolog 2 (SNAI2), twist homolog 1 (TWIST1), and N-cadherin (CDH2) [16] were also analyzed.

Comparative Analysis of Airway BC Signature Gene Expression in Lung AdenoCa and Lung Squamous Cell Carcinoma (SqCa)

To compare the expression of the airway BC signature in lung adenoCa to SqCa, the 58 adenoCa and 53 SqCa described by Bild et al [10] was analyzed. To compare the overall airway BC signature expression between adenoCa and SqCa, the BC index was calculated based on the median level of each airway BC signature gene in the adenoCa subjects, and the I_{BC} values were assessed using Mann-Whitney test. For identification of the airway BC signature genes differentially expressed in BC-high adenoCa *vs* SqCa, the criteria used was $p<0.05$ with a Benjamini-Hochberg correction to limit the false positive rate. Differential expression of selected airway BC signature genes in the analyzed lung cancer subtypes (BC-low adenoCa, BC-high adenoCa and SqCa) was additionally analyzed using Mann-Whitney test.

References

1. Hackett NR, Shaykhiev R, Walters MS, et al. The Human Airway Epithelial Basal Cell Transcriptome. *PLoS One* 2011; 6: e18378.

2. 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.

3. Ding L, Getz G, Wheeler DA, et al. Somatic Mutations Affect Key Pathways in Lung Adenocarcinoma. *Nature* 2008; 455: 1069-1075.

4. 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.

5. 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.

6. 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.

7. 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.

8. Chiang DY, Villanueva A, Hoshida Y, et al. Focal Gains of VEGFA and Molecular Classification of Hepatocellular Carcinoma. *Cancer Res* 2008; 68: 6779-6788.

9. 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.

10. 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.

11. 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.

12. 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.

13. Dvorak A, Tilley AE, Shaykhiev R, et al. Do Airway Epithelium Air-Liquid Cultures Represent the in Vivo Airway Epithelium Transcriptome? *Am J Respir Cell Mol Biol* 2011; 44: 465-473.

14. Zhang L, Whitsett JA, Stripp BR. Regulation of Clara Cell Secretory Protein Gene Transcription by Thyroid Transcription Factor-1. *Biochim Biophys Acta* 1997; 1350: 359-367.

15. Wiedenmann B, Huttner WB. Synaptophysin and Chromogranins/Secretogranins--Widespread Constituents of Distinct Types of Neuroendocrine Vesicles and New Tools in Tumor Diagnosis. *Virchows Arch B Cell Pathol Incl Mol Pathol* 1989; 58: 95-121.

16. Kalluri R, Weinberg RA. The Basics of Epithelial-Mesenchymal Transition. *J Clin Invest* 2009; 119: 1420-1428.

Supplemental Tables

Supplemental Table I. Lung Adenocarcinoma Patient Characteristics

	Primary cohort (Chitalle et al.)	Validation 1 (Bild et al.)	Validation 2 (Shedden et al.)
Number of patients	182	58	327
Age, mean ± S.D.	66.7±10.7	N.A.[2]	64.3±10.2
Gender			
Male	78	N.A.	180
Female	104	N.A.	147
Smoking history			
Never	35	N.A.	28
Ever	146	N.A.	210
Unknown	1	N.A.	89
Pathological stage[1]			
I	63	N.A.	200
II	80	N.A.	65
III	39	N.A.	60
Unknown	0	N.A.	2
Median follow-up (month)	43.8	30.1	48.8
Number of death	90	32	191

[1] Pathological stage was based on 6th edition TNM staging.

[2] Abbreviations: N.A.: not available

Supplemental Table II. Characterization of the 10-Gene Basal Cell (BC)-high Lung Adenocarcinoma (adenoCa) Signature

Gene name	Gene symbol	Functions	Identification of BC-high adenoCa	
			Sensitivity (%)	Specificity (%)
Phosphoglycerate mutase 1 (brain)	PGAM1	Enzyme in the glycolytic pathway	93.5	93.5
Transmembrane protein 158 (gene/pseudogene)	TMEM158	Ras-induced senescence	91.3	82.6
Solute carrier family 16, member 3 (monocarboxylic acid transporters)	SLC16A3	Lactic acid and pyruvate transport	89.1	87
Desmoglein 2	DSG2	Cadherin cell adhesion	89.1	80.4
Serine/threonine kinase receptor-associated protein	STRAP	RNA splicing	89.1	80.4
Glutaredoxin 2	GLRX2	Induction of apoptosis by oxidative stress	87	89.1
ArfGAP with FG repeats 1	AGFG1	RNA trafficking or localization	87	84.8
Leucine rich repeat containing 42	LRRC42	(unknown)	87	84.8
Sec23 homolog A (S. cerevisiae)	SEC23A	Endoplasmic reticulum-Golgi protein trafficking	87	84.8
Procollagen-lysine, 2-oxoglutarate 5-dioxygenase 2	PLOD2	Stability of intermolecular collagen cross-links	87	76.1

1 Based on the microarray analysis of the primary adenoCa data set.

Abbreviations: BC: basal cell, adenoCa: adenocarcinoma, AE: airway epithelium, N.S.: not significant.

Supplemental Table III. Gene Expression Analysis of Airway Basal Cell (BC) Signature in Lung Adenocarcinoma (adenoCa)

Compared gene sets	Cohorts	p value (t-test)	
		< 0.5x median	**> 2x median**
Airway BC signature *vs* non-BC signature	1) Primary cohort (n=182, Chitale et al) 2) Validation cohort 1 (n=58, Bild et al) 3) Validation cohort 2 (n=327, Shedden et al.)	$p<0.002$	$p<0.0006$
Airway BC signature *vs* randomly selected gene sets 1-50	Airway BC signature 1) Primary cohort (n=182, Chitale et al) 2) Validation cohort 1 (n=58, Bild et al) 3) Validation cohort 2 (n=327, Shedden et al.) Ramdomly selected gene sets 1-50 1) Primary cohort (n=182, Chitale et al)	$p<0.0000005$	$p<0.02$

Supplemental Figure Legends

Supplemental Figure 1. Diagram representing experimental flow of the study.

Supplemental Figure 2. Correlation between BC index and NK2 homeobox 1 (NKX2-1) expression in primary lung adenocarcinoma cohort (n=182). Y-axis - BC index; x-axis – normalized NKX2-1 expression levels; Pearson correlation (r) and p value are indicated.

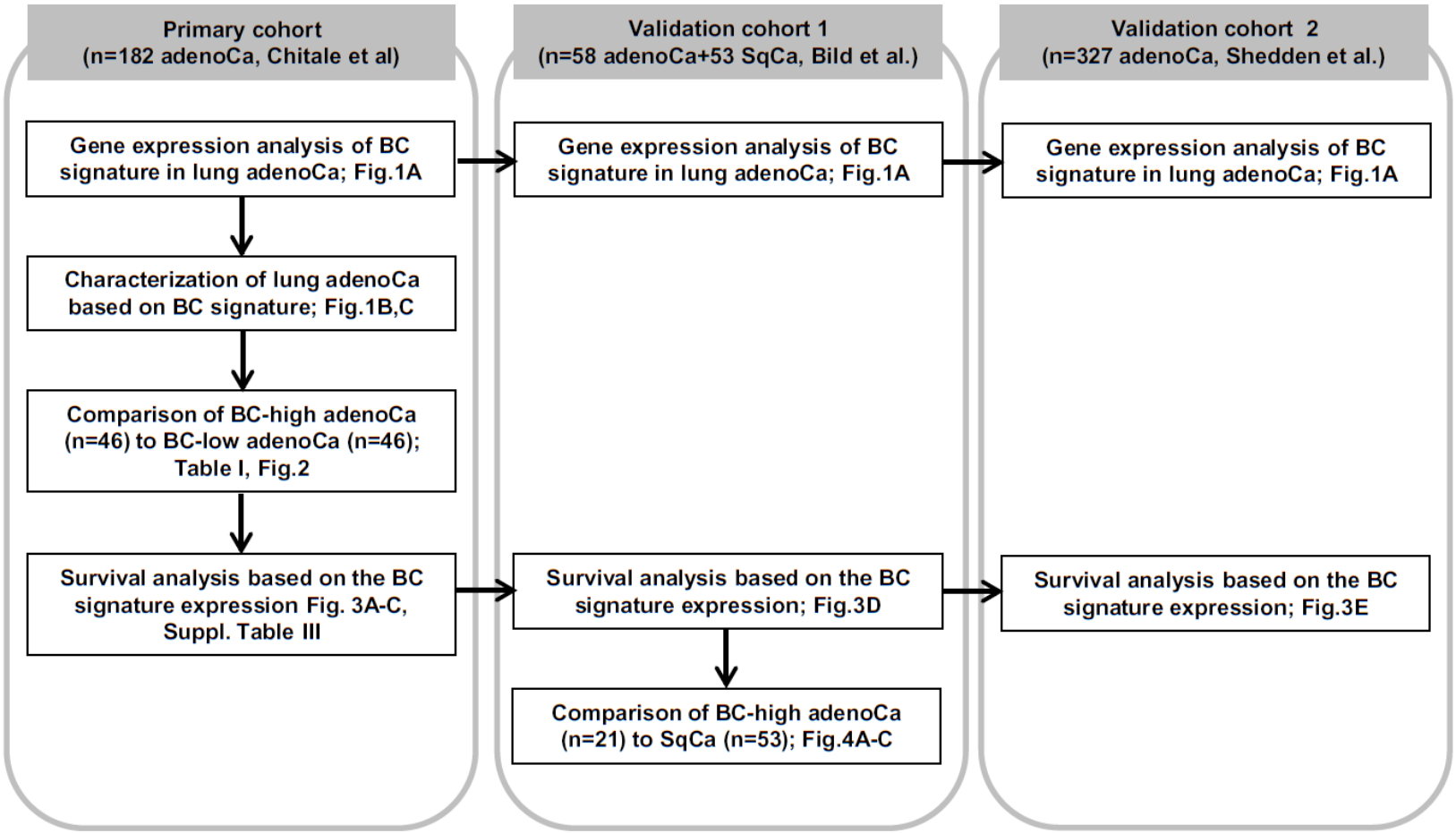
Supplemental Figure 3. Immunohistological analysis of BC-high lung adenoCa, BC-low lung adenoCa, lung squamous cell carcinoma (SqCa) and normal lung tissue for the expression of the thyroid transcription factor-1 (TTF-1) and tumor protein TP63. Representative BC-high and BC-low adenoCa biopsy samples were selected based on the gene expression data as described in Methods; scale bar - 50μm for original magnification x60 for normal lung tissue, 20 μm for original magnification x20 for lung cancer tissue.

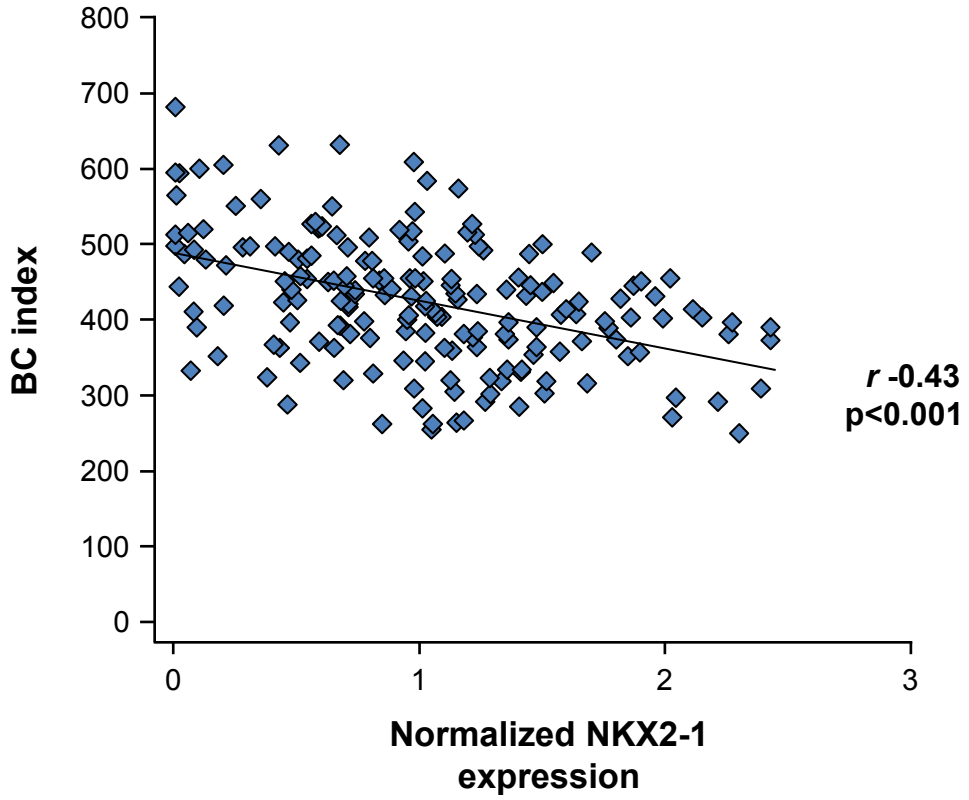
Supplemental Figure 4. Examples of expression of small cell lung carcinoma-related genes in BC-high adenoCa compared to BC-low adenoCa**:** tumor protein p53 (TP53); retinoblastoma 1 (RB1), V-myc myelocytomatosis viral oncogene homolog 1, lung carcinoma derived (avian) (MYCL1, also known as L-MYC). In all panels, log2-transformed normalized 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); ◦ -1.5 x IQR to 3 x IQR, * - more or less than 3 x IQR.

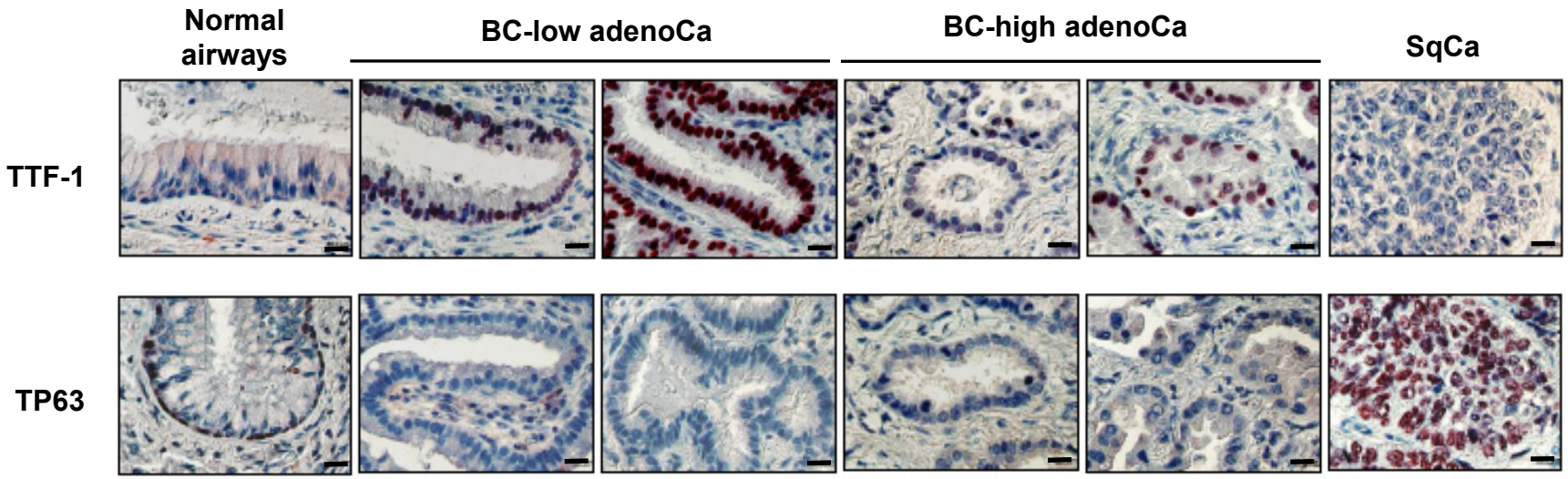
Supplemental Figure 5. Disease free survival after surgery of BC-high adenoCa (red) vs BC-low adenoCa (blue) in primary lung adenocarcinoma cohort (n=182). p values were determined by the log-rank test; the number of individuals in each group are indicated.

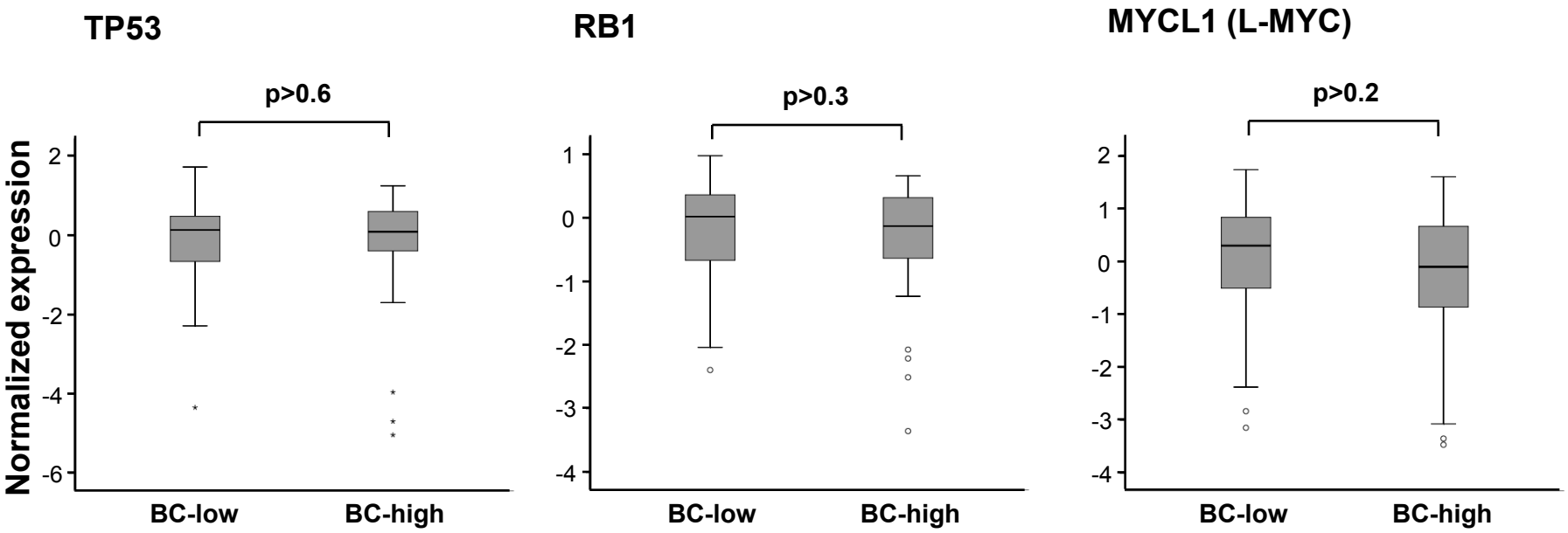
Supplemental Figure 6. Survival analysis of lung squamous cell carcinoma (SqCa) patients based on the BC signature expression in the lung cancer cohort (Bild et al.) compared to the adenoCa. **A.** Categorization of BC-high (red) and BC-low (blue) patients with adenocarcinoma (adenoCa) and SqCa. SqCa samples were categorized based on the adenoCa cohort, in which

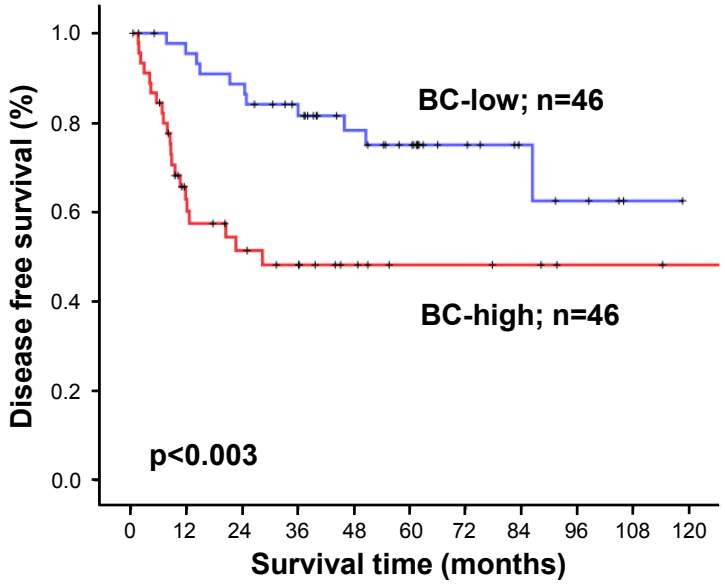
BC-high and BC-low samples were determined as described in Methods. **B.** Overall survival of BC-high SqCa (red) *vs* BC-low SqCa patients (blue) **C.** Overall survival of BC-high adenoCa (red), BC-low adenoCa and SqCa. p values were determined by the log-rank test; the number of individuals in each group are indicated.

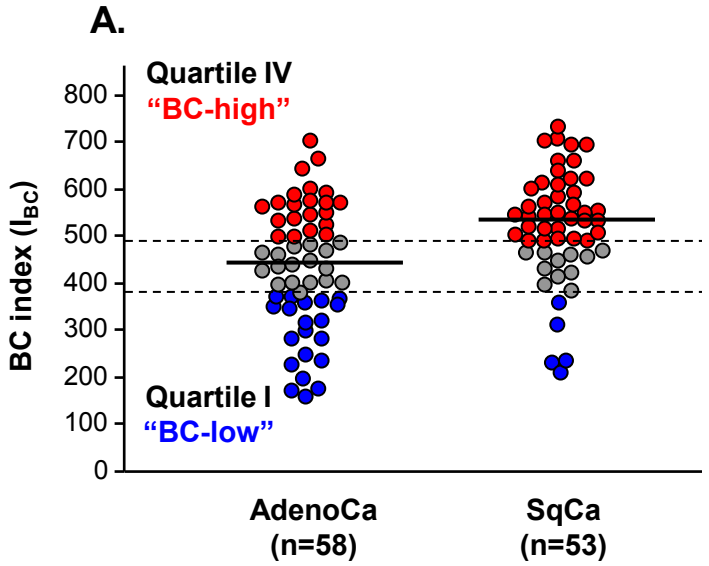

Primary cohort
(n=182 adenoCa, Chitale et al)
Validation cohort 1
(n=58 adenoCa+53 SqCa, Bild et al.)
Validation cohort 2
(n=327 adenoCa, Shedden et al.)
Gene expression analysis of BC signature in lung adenoCa; Fig.1A
Gene expression analysis of BC signature in lung adenoCa; Fig.1A
Gene expression analysis of BC signature in lung adenoCa; Fig.1A
Characterization of lung adenoCa based on BC signature; Fig.1B,C
Comparison of BC-high adenoCa (n=46) to BC-low adenoCa (n=46); Table I, Fig.2
Survival analysis based on the BC signature expression Fig. 3A-C, Suppl. Table III
Survival analysis based on the BC signature expression; Fig.3D
Survival analysis based on the BC signature expression; Fig.3E
Comparison of BC-high adenoCa (n=21) to SqCa (n=53); Fig.4A-C

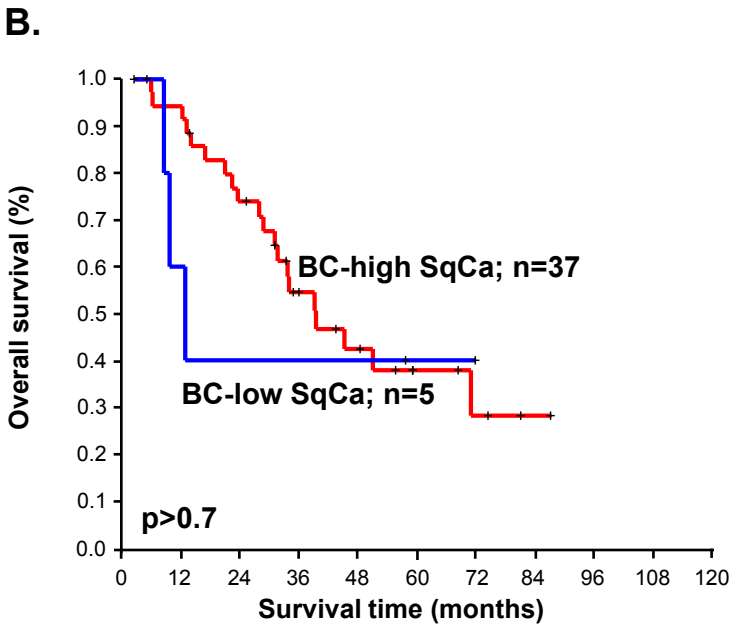

800
700
600
500
400
300
200
100
0
BC index
0
1
2
3
Normalized NKX2-1 expression
r -0.43
p<0.001

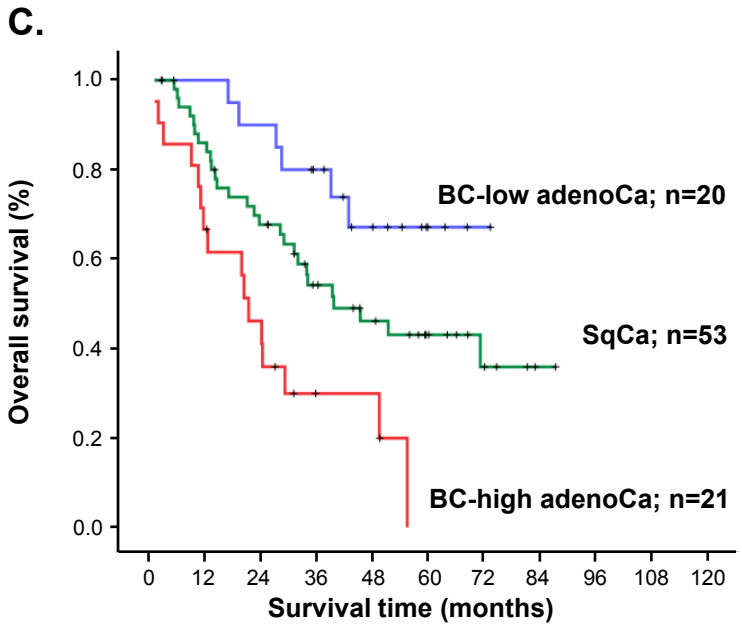

Normal airways
BC-low adenoCa
BC-high adenoCa
SqCa
TTF-1
TP63


TP53
p>0.6
Normalized expression
2
0
-2
-4
-6
BC-low
BC-high
RB1
p>0.3
1
0
-1
-2
-3
-4
BC-low
BC-high
MYCL1 (L-MYC)
p>0.2
2
1
0
-1
-2
-3
-4
BC-low
BC-high


Disease free survival (%)
1.0
0.8
0.6
0.4
0.2
0.0
BC-low; n=46
BC-high; n=46
p<0.003
0
12
24
36
48
60
72
84
96
108
120
Survival time (months)







BC-high adenoCa vs BC-low adenoCa; p<0.001
BC-high adenoCa vs SqCa; p<0.008
BC-low adenoCa vs SqCa; p>0.06