Clinical Cancer Research

Lung Squamous Cell Carcinoma mRNA Expression Subtypes Are Reproducible, Clinically Important, and Correspond to Normal Cell Types

Matthew D. Wilkerson[1], Xiaoying Yin[1], Katherine A. Hoadley[1,2], Yufeng Liu[3,4], Michele C. Hayward[1], Christopher R. Cabanski[3], Kenneth Muldrew[5], C. Ryan Miller[1,5], Scott H. Randell[1,6], Mark A. Socinski[1,7], Alden M. Parsons[7], William K. Funkhouser[1,5], Carrie B. Lee[1,7], Patrick J. Roberts[1], Leigh Thorne[1,5], Philip S. Bernard[8], Charles M. Perou[1,2], and D. Neil Hayes[1,7]

Abstract

Purpose: Lung squamous cell carcinoma (SCC) is clinically and genetically heterogeneous, and current diagnostic practices do not adequately substratify this heterogeneity. A robust, biologically based SCC subclassification may describe this variability and lead to more precise patient prognosis and management. We sought to determine if SCC mRNA expression subtypes exist, are reproducible across multiple patient cohorts, and are clinically relevant.

Experimental Design: Subtypes were detected by unsupervised consensus clustering in five published discovery cohorts of mRNA microarrays, totaling 382 SCC patients. An independent validation cohort of 56 SCC patients was collected and assayed by microarrays. A nearest-centroid subtype predictor was built using discovery cohorts. Validation cohort subtypes were predicted and evaluated for confirmation. Subtype survival outcome, clinical covariates, and biological processes were compared by statistical and bioinformatic methods.

Results: Four lung SCC mRNA expression subtypes, named primitive, classical, secretory, and basal, were detected and independently validated ($P < 0.001$). The primitive subtype had the worst survival outcome ($P < 0.05$) and is an independent predictor of survival ($P < 0.05$). Tumor differentiation and patient sex were associated with subtype. The expression profiles of the subtypes contained distinct biological processes (primitive: proliferation; classical: xenobiotic metabolism; secretory: immune response; basal: cell adhesion) and suggested distinct pharmacologic interventions. Comparison with lung model systems revealed distinct subtype to cell type correspondence.

Conclusions: Lung SCC consists of four mRNA expression subtypes that have different survival outcomes, patient populations, and biological processes. The subtypes stratify patients for more precise prognosis and targeted research. *Clin Cancer Res; 16(19); 4864–75. ©2010 AACR.*

Lung cancer is the leading cause of cancer-related death worldwide (1). Squamous cell carcinoma (SCC) is a major histologic type and comprises ~30% of all pulmonary tumors (2, 3). SCC is defined by the presence of cytoplasmic keratinization and/or desmosomes (intracellular bridges; ref. 4). Clinically, SCC tumors occur more often in smokers and males compared with the other histologic types (2, 5). Patients affected with SCC tumors show a wide range of clinical outcomes. For instance, 83% of autopsied SCC patients had regional metastases (5) and 68% of SCC stage I patients survived beyond 5 years (6). Within SCC, there is noticeable morphologic variability, especially among poorly differentiated tumors (4, 7). The WHO SCC type includes a stratification of this variability with four variants (papillary, small cell, clear cell, and basaloid; ref. 4), but their prevalence and clinical and biological significance remain unclear. Because there is significant pathologic and clinical outcome variability within the SCC histologic type, a robust, biologically derived subclassification may be valuable.

Recent years have seen progress in classification of a variety of malignancies using full-genome molecular assays, primarily those directed at mRNA expression [e.g., leukemia (8), breast (9), and lung adenocarcinoma (10)]. A

Authors' Affiliations: [1]Lineberger Comprehensive Cancer Center, [2]Department of Genetics, [3]Department of Statistics and Operations Research, [4]Carolina Center for Genome Sciences, [5]Department of Pathology and Laboratory Medicine, [6]Department of Cell and Molecular Physiology, and [7]Department of Internal Medicine, Division of Medical Oncology, Multidisciplinary Thoracic Oncology Program, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina and [8]Department of Pathology, University of Utah Health Sciences Center, Salt Lake City, Utah

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

Corresponding Author: D. Neil Hayes, University of North Carolina at Chapel Hill, 450 West Drive, Campus Box 7295, Chapel Hill, NC 27599. Phone: 919-966-3786; Fax: 919-966-1587; E-mail: hayes@med.unc.edu.

doi: 10.1158/1078-0432.CCR-10-0199

Translational Relevance

Lung squamous cell carcinoma (SCC) has broad clinical, genetic, and morphologic heterogeneity. Currently, there is no subclassification that adequately describes this variability and SCC patients are basically treated as though they have the same disease. One explanation for SCC variability is that SCC is not a singular disease but a mixture of multiple discrete diseases or subtypes defined by innate biological differences. Using five discovery cohorts and an independent validation cohort totaling 438 patients, we show that SCC is composed of four robust mRNA expression subtypes (named primitive, classical, secretory, and basal). The subtypes have significantly different survival outcomes, patient populations, and biological processes. Using these subtypes as a basis for a future clinical diagnostic assay, patients could receive a more precise prognosis. Additionally, we described model system partners for the subtypes that can be used for targeted basic research.

successful approach is unsupervised class discovery, which detects naturally occurring tumor classes ("mRNA expression subtypes") without prespecified characteristics such as patient survival (8). Preliminary efforts have been made in SCC, suggesting the existence of SCC mRNA expression subtypes. In independent analyses, investigators (11–13) discovered two mRNA expression subtypes with intriguing biological profiles and a corresponding patient survival difference. These studies show that SCC might be subclassified using mRNA expression into groups with clinical relevance; however, the studies were not done in a manner that validated either the number or the nature of these intriguing classes. A validated mRNA expression classification could substantially progress patient care and research in lung SCC. In this study, we describe four novel reproducible expression subtypes (primitive, classical, secretory, and basal) of lung SCC. The SCC subtypes have different survival outcomes, patient demographics, physical characteristics, biological processes, and correspondence to normal lung cell types.

Materials and Methods

Tumor collection

Frozen, surgically extracted, macrodissected, primary tumors from treatment-naive patients at the University of North Carolina with a lung SCC diagnosis were collected under Institutional Review Board approved protocols 90-0573 and 07-0120. Morphologic quality control was based on a review of a representative H&E-stained section from paraffin-embedded tissue immediately adjacent to the frozen tissue for confirmation of squamous histology by four pathologists (Supplementary Fig. S1) and for quantification of tumor content. Tumor RNA was extracted (14) and assayed for mRNA expression using Agilent 44,000 probe microarrays for a total of 56 microarrays. Microarrays were processed by normexp background correction and loess normalization (15). This data set is referred to as the "validation cohort" and was deposited at National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE17710).

Published data sets

A structured search for publicly available SCC mRNA expression microarray data sets was conducted via Gene Expression Omnibus and PubMed and manually selecting data sets that have a large number of lung SCC samples to permit subtype analysis and that have significant cross–data set gene reliability, as measured by integrative correlations (16). This search yielded five data sets (referred to as the "discovery cohorts") from the following studies: Bild et al. (17), Expression Project for Oncology (Expo; http://www.intgen.org/expo/), Lee et al. (18), Raponi et al. (13), and Roepman et al. (19). Published cohorts contained surgical resections from treatment-naive patients if indicated. Clinical data and raw or processed microarray data were obtained. Only microarrays with SCC histology were retained. Raw microarrays or gene lists from lung model systems were obtained (20–23). Microarrays were subjected to standard quality assessments, mapped to a common transcript database, and processed into gene-level expression values (Supplementary Table S1).

Unsupervised subtype discovery

The subtype discovery and validation procedure is depicted in a flowchart (Supplementary Fig. S2). Genes with high reliability and variability were selected similar to previously described methods (9, 10, 12, 13, 16). Gene reliability was measured by integrative correlations, and genes having an estimated false discovery rate (FDR) of 0.1% were retained (16). To select variable genes, genes in each discovery cohort were ranked by median absolute deviation in decreasing order. These ranks were averaged and reranked to make a single, ranked gene list. The top 25% of this ranked list, totaling 2,307 genes, was used for clustering. Before clustering, each data set was gene median centered (24, 25). Subtypes were determined in each discovery cohort by the Consensus Clustering algorithm via ConsensusClusterPlus (26, 27). This algorithm completed 1,000 microarray subsamples at a proportion of 80% and clustered these subsamples by an agglomerative average-linkage hierarchical algorithm using 1-Pearson correlation coefficient distance. Consensus values, the proportion that two microarrays occupy the same cluster, were calculated and then clustered by an agglomerative average-linkage hierarchical algorithm using Euclidean distance.

Subtype summarization by centroids

Centroids are median expression profiles of a group of arrays and were prepared using methods previously described (25, 28). Centroids were determined by taking a group of microarrays from a gene median centered cohort and obtaining the median of each gene. Multicohort centroids are determined by taking a group of centroids and taking the median of each gene.

Differentially expressed genes

Differentially expressed genes were determined by a standardized mean difference procedure that considers between cohort and within-cohort variation (29) using the GeneMeta Bioconductor library (http://bioconductor.org/packages/2.2/bioc/html/GeneMeta.html) and a random effects option. Gene set enrichment analysis (GSEA) was used to determine gene sets significantly enriched in ranked gene lists (30).

Validation cohort subtype prediction

Subtype status of the validation cohort was predicted by a nearest-centroid classification algorithm following previously published methods (28). In brief, the predictor was built, using only the discovery cohorts, by adding genes to a balanced centroid, assessing subtype prediction error rates by leave-one-out cross-validation, adding genes differentially expressed from the most mispredicted subtype to its centroid, and stopping once accuracy failed to improve. Subtype predictor centroids, unsupervised gene lists, and all gene multicohort centroids are available online (http://cancer.unc.edu/nhayes/publications/scc/).

Survival analysis

The R library survival was used for survival statistical analyses. Patients dead within 1 month following surgery were considered to have procedure-related complications and not considered in survival analyses. Five patients met this condition all from the UNC cohort. Relapse-free survival (RFS) time was defined as the time from surgery until first relapse or death.

Immunohistochemistry

Cores (1 mm) were taken from available UNC cohort tissue blocks and randomly organized into tissue microarray blocks. Consequal 4-μm array block sections were assembled on array slides and stained with H&E, MAC387 (Dako), p63 (Dako), CK7 (Leica Microsystems), and MCM6 (Santa Cruz Biotechnology).

Computational procedures were executed using R version 2.7.1 (http://www.r-project.org/) and Bioconductor libraries (http://www.bioconductor.org) unless otherwise specified.

Results

Unsupervised discovery of lung SCC expression subtypes in five cohorts

Lung SCCs are a heterogeneous group of tumors, and therefore, we did a common set of mRNA expression analyses using five previously published lung SCC data sets to determine how many distinct subtypes/groups of disease might exist. These five discovery cohorts were analyzed for the presence of mRNA expression subtypes using the Consensus Clustering methodology (26) as previously described for lung cancer (10). Consensus Clustering is a semiquantitative method for determining an optimal number of mRNA expression clusters/groups. Results show that all five cohorts contain four clusters (Supplementary Fig. S3). There is no compelling evidence for a higher number of clusters. To test if the four clusters from each cohort have the same expression profiles, a published centroid clustering method was followed (10). The centroid clustering shows a four-group structure, where each cohort is in each group, with only one cohort absent in one group (Supplementary Fig. S4). Therefore, the four clusters (mRNA expression subtypes) found in the five discovery cohorts have consistent expression profiles. To derive the optimal subtype for each patient, a multicohort centroid classification was used to assign each patient to a subtype, similar to published methods (28). A centroid clustering based on these optimal subtypes again shows a four-group structure and complete, unambiguous cross-cohort correspondence (Fig. 1). The cross-cohort clustering is statistically significant [Sigclust (31) *P* values in Fig. 1]. Interestingly, the subtypes have approximately the same prevalence among the discovery cohorts (Table 1). Using biological characteristics described below, the lung SCC mRNA expression subtypes are named primitive, classical, secretory, and basal.

SCC subtype independent validation

Although the four SCC subtypes were "cross-cohort" validated in that they were repeatedly found in five cohorts, this validation was not independent because discovery co-occurred with validation. For an independent validation, we tested the hypothesis that the SCC subtypes will exist in a new discovery-independent cohort. To test this hypothesis, a subtype predictor was built using the discovery cohorts, which consisted of 208 genes and had 94% leave-one-out cross-validation accuracy. Using this predictor, subtype classifications were made for microarrays from a new cohort of 56 lung SCC tumors collected at UNC. All four subtypes were predicted in the UNC cohort and in approximately the same prevalence as the discovery cohorts (Fig. 2; Table 1), which supports subtype reproducibility. To confirm the validity of the predictions, a comparison of expression characteristics between the discovery and UNC cohorts was completed similar to a recent related study (32). We compiled a large validation gene set of the top 100 genes overexpressed and underexpressed per subtype of the discovery cohorts (Fig. 2A), which yielded 1,117 unique genes. Subtype expression patterns are highly concordant between the discovery and UNC cohorts across the validation gene set (Fig. 2A and B), confirming that the large-scale expression patterns are consistent beyond the predictor gene set. In addition, the subtypes of

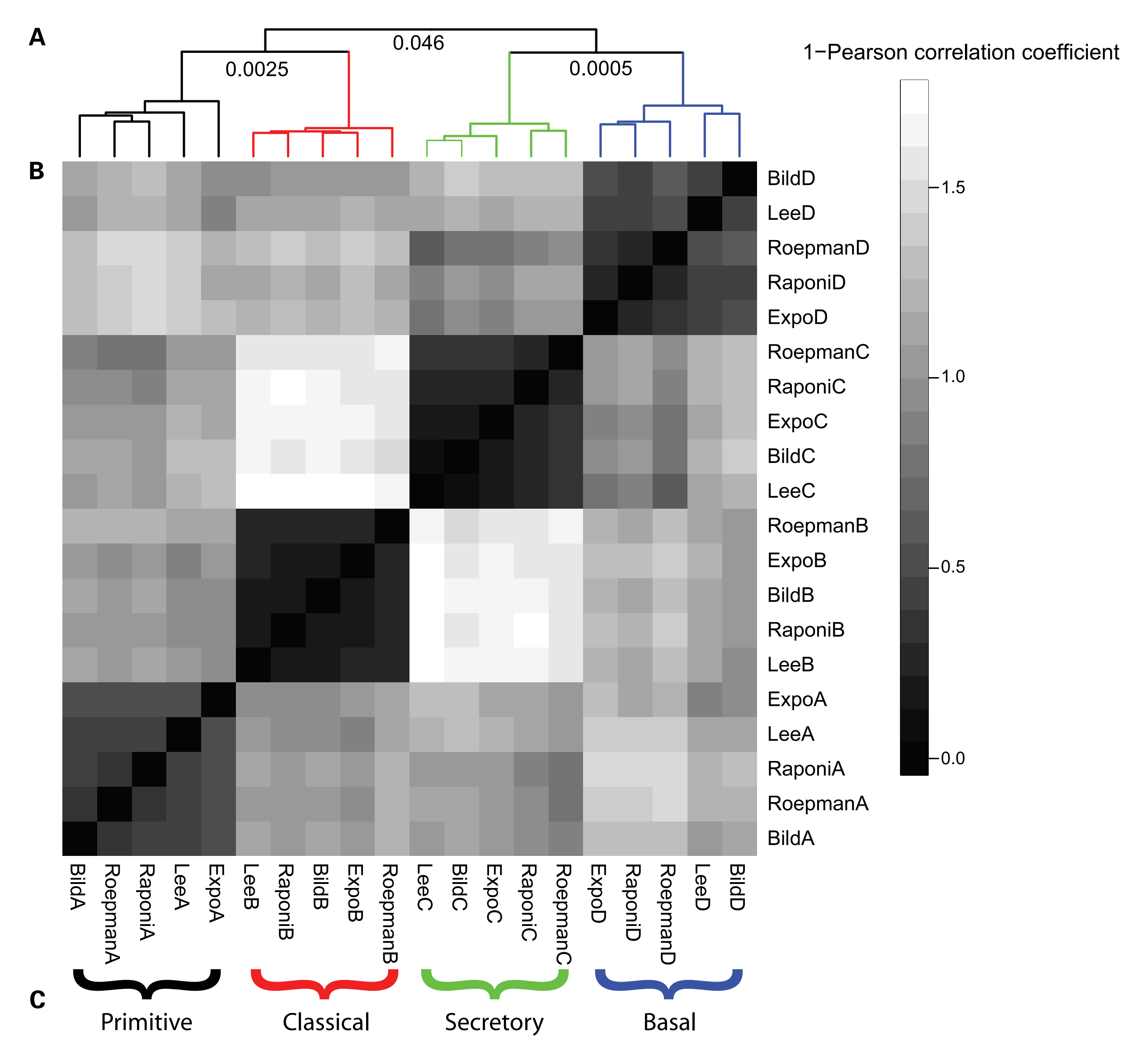


Fig. 1. Discovery cohort correlation matrix and dendrogram. Cells are labeled by discovery cohort and adjusted centroid, where A to D are from Supplementary Fig. S4. B, Cells in the matrix represent the 1-Pearson correlation coefficient between two discovery cohort and adjusted centroids by shading according to the scale above. For example, BildA and RoepmanA have highly similar expression profiles, a large Pearson correlation coefficient, a small 1-Pearson correlation coefficient value, and corresponding cells darkly shaded. A, the matrix is ordered by columns and rows by the dendrogram at the top of the matrix. The dendrogram is the result of an agglomerative, average-linkage, hierarchical clustering using this correlation matrix. C, four expression subtypes. Statistical significance of the three binary divisions leading to the four subtypes is shown by Sigclust (31) *P* values in the dendrogram at the corresponding binary split.

the UNC cohorts are a statistically significant partition of its mRNA expression [SWISSMADE (33) subtypes versus random classes; $P < 0.001$]. We conclude that the predefined SCC subtypes exist in the UNC cohort and are, therefore, independently validated.

To preliminarily evaluate if clinically applicable biomarkers can distinguish the subtypes, we selected one overexpressed gene per subtype (basal, *S100A8*; classical, *TP63*; secretory, *KRT7*; primitive, *MCM6*) for immunohistochemical protein expression comparison using a tissue microarray subset of the UNC cohort ($n = 38$). All antibodies targeting these genes, except *MCM6*, had sufficient staining for analysis. Protein expression clustering using basal, classical, and secretory samples revealed three essentially mutually exclusive groups with one marker defining each group (Supplementary Fig. S5). These groups were significantly associated with tumor subtype ($P = 0.007$, Fisher's exact test). This suggests that SCC subtypes can also be distinguished by immunohistochemistry, and future work may find the optimal panel of immunohistochemical antibodies.

Subtypes exhibit distinct biological processes

To discern biological processes associated with each subtype, subtype mRNA expression was evaluated for enrichment in gene ontology, pathway, transcription factor binding site, and cytoband gene sets by GSEA (30). Because of the inherent redundancy in biology, we have collapsed these processes into functional themes (Fig. 3). Here, subtypes are described in terms of overexpression relative to the other subtypes.

The distinctive functional theme of the primitive subtype is cellular proliferation, which includes genes such as *minichromosome maintenance 10* (*MCM10*), *E2F transcription factor 3* (*E2F3*), *thymidylate synthetase* (*TYMS*), and *polymerase α1* (*POLA1*), and a published proliferation signature (34). This proliferation theme is overexpressed in the most rapidly growing breast cancer cell lines (35) and in the most poorly differentiated, poor survival tumors from various organ sites (34). Complementary to

Table 1. Clinical characteristics of lung SCC expression subtypes

		Discovery cohorts					Validation cohort (UNC)				
		Primitive	Classical	Secretory	Basal	Total	Primitive	Classical	Secretory	Basal	Total
No. patients	Bild et al.	7	20	15	10	52					
	Lee et al.	14	30	22	9	75					
	Expo	4	15	11	6	36					
	Raponi et al.	20	41	32	34	127					
	Roepman et al.	15	35	19	23	92					
	Total	60	141	99	82	382					
							9	21	14	12	56
Age	Median	68	64	66	67	66	65	68	64	72	67
Gender	% Female	36	19	26	29	26	67	33	43	42	43
	% Male	64	81	74	71	74	33	67	57	58	57
Smoking	% Nonsmoker	8	1	3	2	3	0	0	0	0	0
	Mean pack-years	64	72	60	68	66	43	74	62	46	60
Stage	% I	66	58	66	55	61	56	57	57	75	61
	% II	25	26	20	37	26	44	33	36	25	34
	% III	5	17	14	9	13	0	10	7	0	5
	% IV	3	0	0	0	1	0	0	0	0	0
Grade	% Poor	39	15	21	16	21	56	24	43	33	36
	% Moderate	58	82	76	76	75	44	76	57	67	64
	% Well	3	2	3	8	4	0	0	0	0	0
OS	No. patients	42	96	66	67	271	8	19	13	11	51
	% 1-y survival	64	89	84	88	84	88	100	82	90	92
	% 3-y survival	47	63	59	71	62	15	38	48	60	41
% Tumor	Median						90	80	80	93	90
	Interquartile range						75-94	60-100	70-90	73-96	60-95
% Necrosis	Median						5	5	5	4	5
% Fibrosis	Median						15	13	18	10	10
Lymphocytes	% Marked						33	31	40	40	36

NOTE: Percent values indicate the proportion of the samples in a particular subtype with a particular variable (e.g., 36% of the primitive subtype samples came from female patients in the discovery cohort). Some percents may not total 100% due to rounding. OS percents are Kaplan-Meier estimates. Gray shading indicates data unavailability.

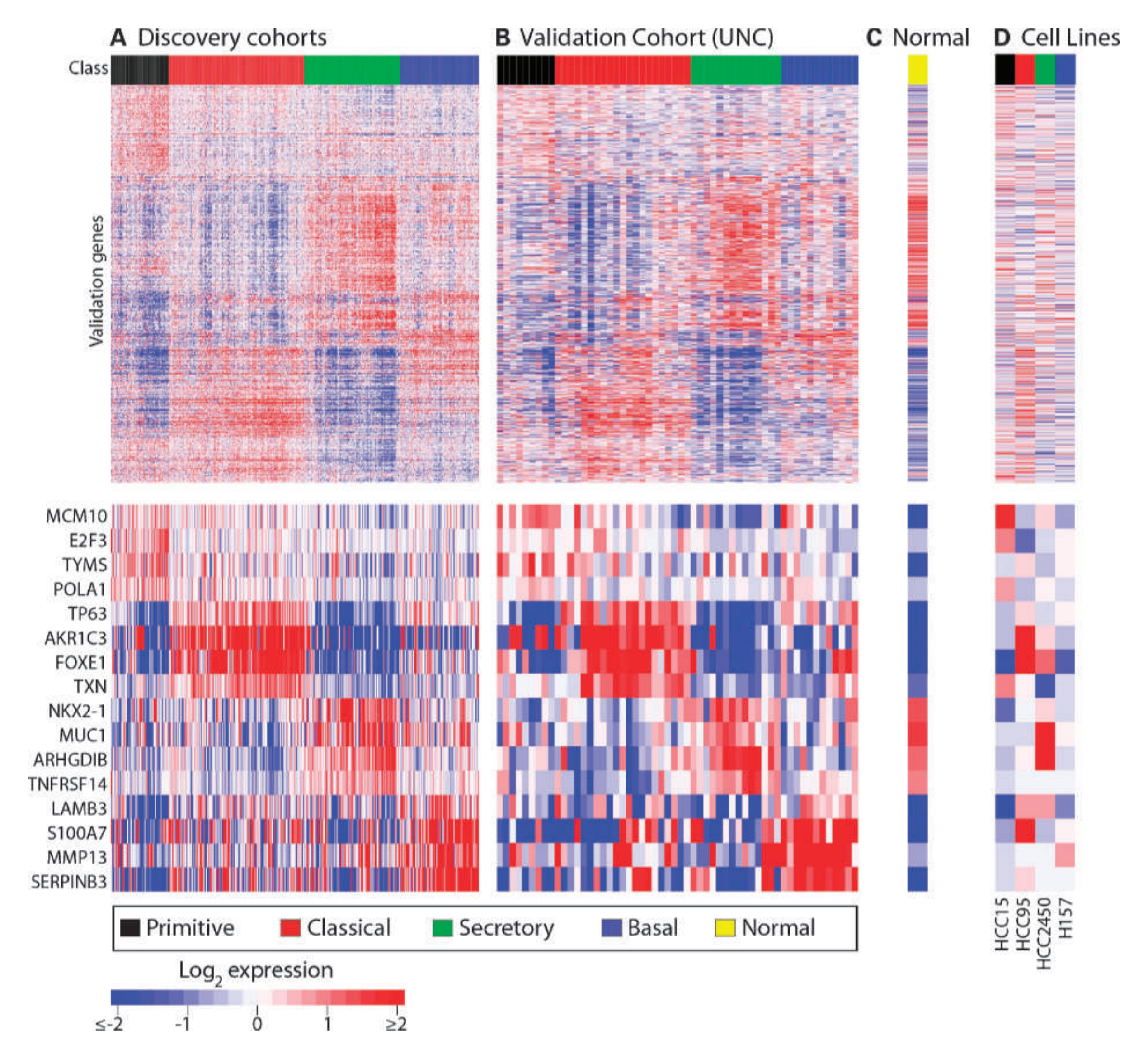


Fig. 2. Independent validation of lung SCC expression subtypes. Heat maps depict mRNA expression of discovery cohorts (A), the validation cohort (B), a normal lung centroid (C), and SCC cell lines (D). Microarrays are columns and are labeled with their class. Genes are rows and are ordered by a discovery cohort hierarchical clustering. The normal lung centroid is scaled to the validation cohort for visualization. Manually selected, lung-relevant, validation genes are displayed separately for viewability.

the cellular proliferation functional theme, target genes of the E2F transcription factor, a known proliferation modulator (36), are overexpressed in this subtype as well as two members of the E2F family, *E2F3* and *E2F8*. Other primitive subtype functional themes are RNA processing and DNA repair, which could be a consequence of the proliferation theme or an independent process.

The classical subtype exhibits the distinctive functional theme of xenobiotic metabolism, which detoxifies foreign chemicals. One study showed overexpression of this theme in smokers' versus nonsmokers' airway transcriptomes, including genes such as *GPX2* and *ALDH3A1* (37). Furthermore, this subtype is enriched with a gene signature derived from lung cell lines exposed to cigarette smoke, including genes such as *AKR1C3* (38). Interestingly, the classical subtype has the greatest concentration of smokers and the heaviest smokers among the subtypes. This theme, including genes such as *GPX2*, *AKR1C1*, *TXNRD1*, and *GSTM3*, was noted as overexpressed in one head and neck SCC subtype (group 4 in ref. 39), suggesting a possible relative to the lung SCC classical subtype. The classical subtype overexpresses *TP63*, a transcription factor essential for stratified squamous epithelium development (40) that is more commonly overexpressed and amplified in lung SCC compared with other histologic types (41). Cytoband gene overexpression, a proxy for underlying genomic DNA amplification, suggests that 3q27-28, which contains *TP63*, is amplified in the classical subtype. The microarrays of this study do not have enough resolution to measure TP63 isoform-specific expression, but this may be a goal of future investigations.

Immune response is the major distinctive functional theme of the secretory subtype and includes genes such as *Rho GDP dissociation inhibitor β* (*ARHGDIB*) and *tumor necrosis factor receptor 14* (*TNFRSF14*). Consistent with this theme, the secretory subtype has a NF-κB regulation theme and NF-κB target gene overexpression. This subtype also overexpresses the lung secretory cell markers mucin (*MUC1*) and pulmonary surfactant proteins (*SFTPC, SFTPB,* and *SFTPD*; refs. 7, 42). Interestingly, *thyroid transcription factor 1* (*NKX2-1/TTF1*), known to be highly expressed in adenocarcinoma (43), is overexpressed in the secretory subtype relative to the other SCC subtypes. This commonality could be a result of the glandular cell structure of adenocarcinoma, which perhaps has secretory properties similar to the SCC secretory subtype. A UNC normal lung centroid shows a very similar expression pattern to the secretory subtype over the independent validation gene list, which was selected without considering normal samples (Fig. 2C). To evaluate any possible difference between the secretory subtype samples and normal samples, an unsupervised clustering was completed using only these microarrays

Primitive subtype
Proliferation: cell cycle, DNA replication, pyrimidine metabolism, purine metabolism, mitosis, cell division, DNA replication, cell cycle
MCM10, MCM3, MCM6, BUB1, TIMELESS, POLA1, TYMS, ATIC, PRIM1, CKAP5, CDK2, E2F3, E2F8, CHEK1
RNA processing: mRNA processing, rRNA processing, tRNA processing, Nuclear mRNA splicing via spliceosome
LSM2, SNRPA, CPSF1, EXOSC5, WDR3, PTBP2, TRMT11
DNA repair: base excision repair, nucleotide excision repair, mismatch repair, DNA repair, response to DNA damage stimulus
LIG1, PARP1, UNG, SSRP1, RECQL4, KIF22, FANCA, BARD1, GTF2H4
Cellular component: nucleoplasm, spliceosome, nucleolus
Transcription factor binding sites: E2F, NRF
Drug targets: TYMS, DNMT1, BCL2, CDK2.
Published signatures: Cellular proliferation (34)

Classical subtype
Energy metabolism: oxidative phosphorylation, citrate cycle, electron transport chain
COX5B, NDUFB5, ATP5G3, COX7B, DLD, SDHD, TXN, ATP6V1F
Xenobiotics metabolism: metabolism of xenobioitics by cytochrome p450, glutathione metabolism
ODC1, GSTA4, GSTM4, GSTO1, GPX2, ALDH3A1, AKR1C3, EPHX1, ADH7, G6PD
Cellular component: mitochondrial inner membrane, respiratory chain
Cytobands. 3q27-28: TP63, BCL6, ABCC5
Published signatures: Lung cell culture 24 hour smoke exposure (38)

Secretory subtype
Immune response: complement and coagulation cascade, antigen processing and presentation, natural killer cell mediated cytotoxicity, leukocyte transendothelial migration, B cell receptor signaling, T cell receptor signaling, toll-like receptor signaling, immune response, inflammatory response, innate immune response, cellular defense response, defense response, humoral immune response, T cell activation
SERPINA1, C2, C5, VAV1, GZMB, ITGAM, NFKBIE, ARHGDIB, TNFRSF14, HLA-DPA1, IL32, ALOX5, AIF1, DPP4, TCIRG1, TLR2, TLR4, IRF7
Positive regulation of I-kappaB kinase/NF-kappaB cascade:
RIPK2, CASP1, RHOA, TGM2, MYD88, APOL3,
Transcription factor binding sites: PEA3, NFKB, AML, IRF1
Cellular component: external side of plasma membrane, lysosome
Drug targets: C1QA, CSF2RA, CSF2RB, IL3RA, ALOX5, DPYD, SOAT1, TCN2

Basal subtype
Cell adhesion: ECM recepor interaction, focal adhesion, cell adhesion, cell matrix adhesion, homophillic adhesion
ITGB4, LAMB3, COL11A1, COL17A1, LAMC2, RAC1, ACTN1, PGF, ITGB5, TNFAIP6, CLDN1, HES1, CLSTN1
Epidermal development: epidermis development, keratinocyte differentiation
GJB5, S100A7, KRT5, FABP5, COL17A1, LAMC2
Cellular component: proteinaceous extracellular membrane, basement membrane, collagen
Drug targets: TCN1, MMP3

Fig. 3. Subtype biological functional themes. Significantly enriched gene sets that are overexpressed in a subtype (GSEA preranked, FDR < 0.05) and genes representative of the set are shown. Pathways and biological processes are organized into functional themes, indicated by italics. Transcription factor binding site refers to gene sets having a predicted transcription factor binding site. Cellular component refers to gene sets having a particular cellular location. Drug targets are defined as overexpressed in all pairwise subtype comparisons (FDR < 0.01).

(Supplementary Fig. S6). Secretory and normal microarrays clustered with their group in essentially all cases, suggesting that the secretory subtype and normal lung are distinct mRNA expression groups.

The basal subtype expression profile shows a cell adhesion functional theme, including genes such as the laminins (*LAMB3* and *LAMC2*), collagens (*COL11A1* and *COL17A1*), integrins (*ITGB4* and *ITGB5*), and *claudin 1* (*CLDN1*). Additionally, this subtype has an epidermal development theme, including *keratin 5* (*KRT5*), *psoriasin* (*S100A7*), and *gap junction protein β5* (*GJB5*). Several of the genes of the basal subtype, such as *COL17A1*, *LAMC2*, and *CDH3*, are common with a head and neck SCC subtype (group 1 in ref. 39) and a breast cancer subtype (basal-like in ref. 9), suggesting that these different organ site subtypes may share biological properties. The basal subtype overexpresses several S100 family genes: *S100A2*, *S100A3*, *S100A7*, *S100A8*, *S100A9*, *S100A12*, and *S100A14*. *S100A8* and *S100A9* are highly expressed in the basal layer in psoriatic epidermal tissue (44). *S100A2* is a marker specific for the basal layer of the lung epithelium and SCC (45). *KRT5* is a basal layer marker in epithelial tissue (46). The basal subtype is enriched with genes whose products are localized in the basement membrane.

In parallel to differential biological functions are patterns of mRNA expression with implications for pharmacologic intervention (Fig. 3). For example, *TYMS*, a target of antifolates including pemetrexed, is overexpressed in the primitive subtype. The antifolate metabolism pathway is differentially expressed among SCC subtypes, with the secretory subtype showing underexpression and similarity to adenocarcinoma (Supplementary Fig. S7). Overexpression of *TYMS* has been shown to be related to pemetrexed resistance in a dose-dependent manner in lung cancer cell culture (47). In addition, *PARP1*, a target of several drugs in development, is overexpressed in the primitive subtype.

SCC subtype tumor morphologic and patient characteristics

The morphologic and patient characteristics of the subtypes are displayed in Table 1. Grade is significantly associated with subtype (*P* = 0.024, Fisher's exact test). The primitive subtype has an overrepresentation of poorly differentiated tumors, and the basal subtype has an overrepresentation of well-differentiated tumors. Tumor stage is not appreciably different among subtypes, although we note that the classical and secretory subtypes have increased proportions of stage III tumors. The surgical cohorts oversample early stages, and possibly, greater sampling of late-stage patients may find additional subtype-stage associations. Specimen quality metrics of percent tumor, percent necrosis, and percent lymphocyte infiltration are not appreciably different among the subtypes, arguing against sampling artifacts as the source of the subtypes. Two cases of WHO morphologic SCC subclass were definitively called by pathologist review (one basaloid in primitive and classical subtypes), suggesting that these SCC morphologic subclasses are rare.

Patient sex approaches statistically significant association with subtype (*P* = 0.058, Fisher's exact test). Females are overrepresented in the primitive subtype and males in the classical subtype. Consistent with the smoking expression profile of the classical subtype, the classical subtype has the greatest mean pack-years (73; *P* = 0.319, Kruskal-Wallis test) and the lowest proportion of nonsmokers (1%; *P* = 0.214, Fisher's exact test), although these observations do not meet statistical significance.

SCC subtypes have different patient survival outcomes

Overall survival (OS) and RFS outcomes are significantly different among SCC subtypes (Fig. 4). The primitive subtype has worse OS and RFS compared with the other subtypes in all stages and in stage I (Fig. 4), whereas the basal, secretory, and classical seem to have similar outcomes. Considering the UNC cohort alone, the primitive subtype outcome is also worse compared with the other subtypes over all stages (OS: *P* = 0.066, log-rank test; RFS: *P* = 0.004, log-rank test) and stage I (OS: *P* = 0.057, log-rank test; RFS: *P* = 0.007, log-rank test). In the UNC cohort, 7 of 18 recurrences were extrapulmonary and the basal subtype had the lowest number and proportion (0/3). To evaluate the independent contribution of SCC subtype to patient risk in light of known prognostic factors, univariate and multivariate Cox proportional hazard models were constructed (Supplementary Table S2). Significant univariate predictors were primitive subtype for OS and RFS and tumor stage for OS. Patient age and tumor grade were not significant predictors of either outcome. In multiple variable models, only subtype retained significance for OS and RFS. The nonsignificant prediction of the tumor stage may be due to the underrepresentation of late-stage patients across the cohorts.

Raponi et al. reported two SCC mRNA expression subtypes with a survival difference and provided a list of differentially expressed genes, where high expression of the "majority of the genes were downregulated in the high-risk group" (13). Comparison of Raponi et al.'s microarrays by their gene list and the subtypes discovered in this study shows two clear subtype groups: underexpression (primitive and secretory) and overexpression (basal and classical; Supplementary Fig. S8). Therefore, the four subtypes discovered in this study map to prior results and this study has divided each of the prior subtypes into two new ones and improves the SCC mRNA expression subtype granularity. Interestingly, the Raponi et al. poor survival subtype totals 43% of their patients, where the poor survival subtype of this study (primitive) is 16% of their patients. It seems that a fraction of Raponi et al.'s high-risk subtype shows poor survival outcome relative to the remainder of SCC.

SCC subtypes are similar to different normal lung cell types and SCC cell lines

To evaluate the hypothesis that SCC subtypes are derived from different cell types present in the normal lung,

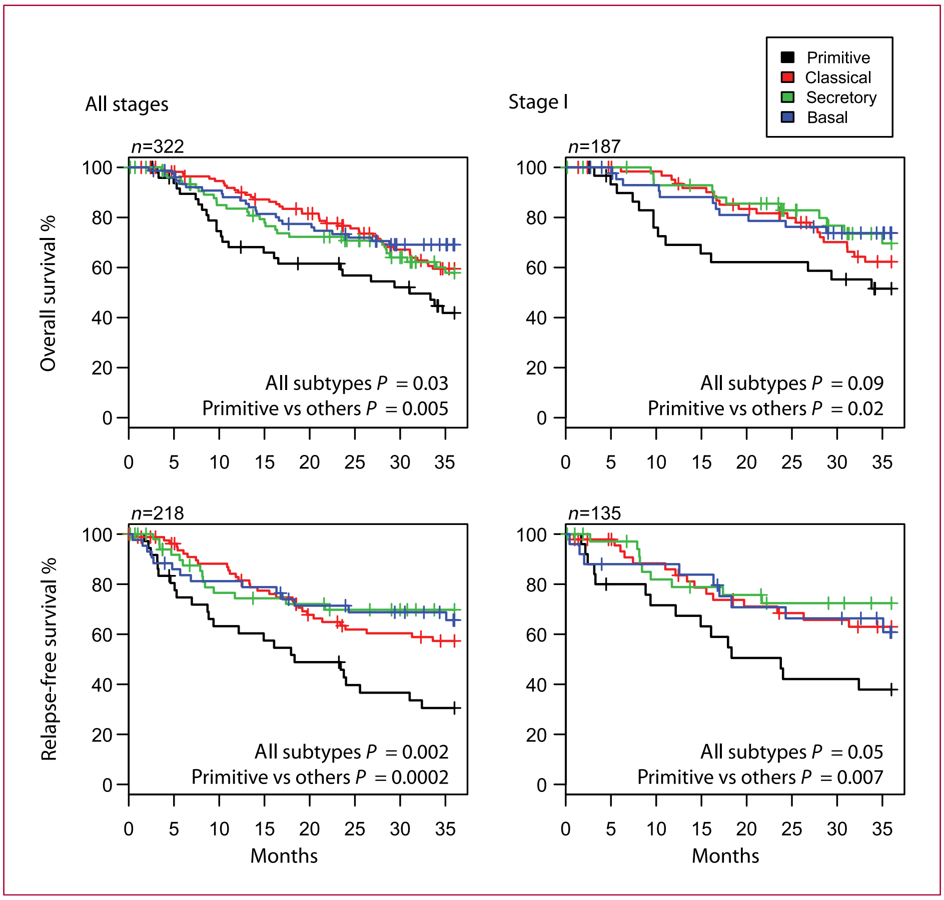


Fig. 4. Survival outcomes of SCC subtypes. Survival was estimated by the Kaplan-Meier method using the available data of all cohorts. The sample sizes (*n*) are different than the overall study sample size due to data availability (OS: Bild et al., Raponi et al., Roepman et al., and UNC cohorts; RFS: Lee et al., Roepman et al., and UNC cohorts). *P* values are from log-rank tests evaluating the independence of survival and subtype.

SCC subtypes were compared by mRNA expression to three published model systems. The first model, "Mouse lung development," is a time series of mouse lungs extracted from embryonic stages to adult (21). Expression similarity is defined as high positive Pearson correlation between an SCC subtype and time points within the model. The primitive subtype shows expression similarity to early-stage mouse lung, and the secretory subtype shows similarity to late-stage mouse lung (Fig. 5A). The second model, "Human bronchial epithelial cell air liquid interface culture" (HBEC-ALIC), is a time series of cultured normal, healthy, human bronchial epithelial cells, in which the early time points consist of stratified basal cells and later time points include secretory and ciliated cells (22). The basal subtype showed expression similarity to the early time points during which basal cells are predominant (Fig. 5B). The primitive and secretory subtypes show expression similarity to the later time points at which there are secretory and ciliated cells. The third model system, "Human microdissected lung cell compartments" (HMLCC), was laser capture microdissected cells contained in surface epithelium and in submucosal glands of normal healthy lung (20). The secretory subtype overexpresses genes that are overexpressed in submucosal glands (Fig. 5C). The basal subtype overexpresses genes that are overexpressed in surface epithelia. The classical subtype does not show appreciable similarity to any specific lung model, is the only subtype to have this property, and could be most similar to multiple or unobserved cell types. Therefore, by the combination of all three lung models, three of the four SCC subtypes have unique similarities to different, normal lung cell types.

In addition to the cell type models, SCC subtypes may correspond to different SCC cell lines, which could establish additional manipulatable models for future investigations into subtype biology. To ascertain if SCC cell lines correspond to different SCC tumor subtypes by mRNA expression, four published SCC cell line microarrays (23) were given subtype classifications by the nearest-centroid predictor. Interestingly, the four cell lines were predicted to be different subtypes (Fig. 2D). Expression of the subtypes between the cell lines and tumors is consistent over the validation gene set (Fig. 2A and D). For example, genes are consistent and mutually exclusive in the cell lines as predicted (HCC15, primitive and *MCM10*; HCC95, classical and *AKR1C3*; HCC2450, secretory and *MUC1*; H157, basal and *MMP13*).

Discussion

The principal novel hypothesis tested in this study is that lung SCC expression subtypes exist, are reproducible, are clinically relevant, and exhibit patterns that correlate with unique cell types in the normal lung. These subtypes (primitive, basal, secretory, and classical) were identified in an unbiased and objective manner and are supported by cross-cohort validation using five training cohorts and by independent validation using a sixth cohort, which together total 438 patients. The expression subtypes were also found in a wide variety of patient populations from the United States, Asia, and Europe, in a wide variety of cohort sizes from 36 to 127. All cohorts showed approximately the same subtype proportions (overall: primitive, 16%; classical, 37%; secretory, 26%; basal, 21%). These subtypes were associated with tumor differentiation and patient sex. Survival outcomes are significantly different among the subtypes, and subtype is an independent predictor of survival. Possible limitations of our analysis include possible sample quality artifacts or patient behavior, such as smoking immediately before surgery; however, all six cohorts showed the same results, so any limitation would have to occur in six large, independently collected cohorts.

The SCC expression subtypes are biologically distinct and show similarities to distinct normal lung cell populations. These biological characteristics serve as the basis for the SCC nomenclature. The basal subtype exhibits many characteristics of lung basal cells, such as cell adhesion and epidermal development functional themes, *S100A2* and *KRT5* basal cell markers, overexpression of genes whose products are localized in the basement membrane, similarity to basal cells in the HBEC-ALIC model, and similarity to surface epithelia in the HMLCC model. The secretory subtype has many features of lung secretory cells, such as surfactant and mucin overexpression, similarity to secretory cells in the HBEC-ALIC model, and similarity to submucosal glands in the HMLCC model. The primitive subtype has a cellular proliferation functional theme, the worst survival outcome, an overabundance of female patients, the most nonsmokers, and an overabundance of poorly differentiated tumors. This subtype is similar to early embryonic mouse lungs, where primitive, less differentiated cells may be predominant and would be consistent with the poorly differentiated nature of these tumors. The primitive subtype also has similarity to late-stage HBEC-ALIC, which could be explained by lung "transient expression" in which differentiation markers are expressed during early lung formation and again in the developed lung (48). Alternatively, a late-emerging and late-active cell type in HBEC-ALIC may be most similar to the embryonic mouse lung. The classical subtype exhibits features representative of typical lung SCC, including the highest prevalence at 37%, overabundance of males, greatest patient smoking behavior, overexpression of *TP63*, and putative amplification of the *TP63*-containing locus 3q27-28.

The distinct SCC subtype to cell population similarities could be explained by the SCC subtypes having different ancestor cells. These different ancestor cells could be cell types of distinct lineages or cellular differentiation stages such as proposed in breast cancer (49). This scenario provides a reason why the SCC subtypes have dramatically different mRNA expression. The subtypes could arise by genetic mutation from different ancestors

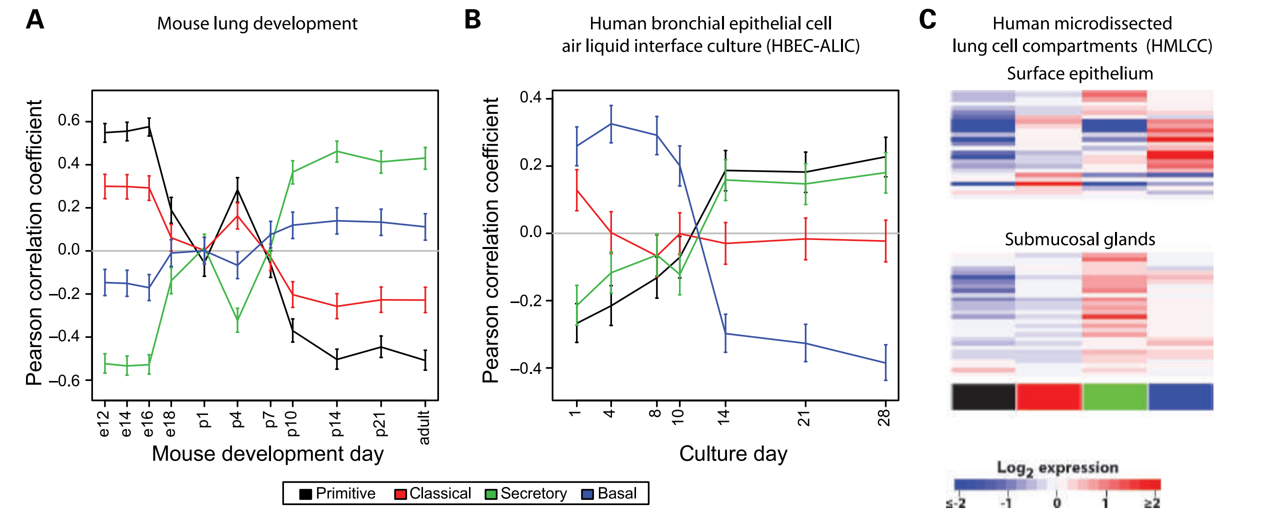


Fig. 5. SCC subtypes compared with lung cell type models. The relationship of relative SCC subtype expression differences to relative expression differences of published lung model systems. A and B, the models Mouse lung development (21) and HBEC-ALIC (22) are microarray time series, where time is indicated on the horizontal axis. Points mark Pearson correlation coefficients of SCC subtype centroids to model time points using the top 1,000 genes having the greatest Pearson correlation coefficient with time. Bars represent 95% confidence intervals. Lines connect points corresponding to the same subtype. Large positive correlations indicate mRNA expression similarity, whereas large negative correlations indicate dissimilarity.
In A, "e" refers to embryonic day and "p" refers to postnatal day. C, the model HMLCC (20) is compared with SCC subtypes via a heat map of genes that are overexpressed in submucosal glands and in surface epithelium as rows and subtype centroids in columns.

that have different mRNA expression, and this ancestral mRNA expression could persist in progeny tumor cells. This putative subtype ancestral cell information could be used in developing SCC subtype pharmacologic interventions that exploit differences in the ancestral cell types. A caveat to our interpretation of SCC subtype to cell population similarity is that the similarity could be caused by coincidence and expression similarities could reflect similar biology and not similar origin. The lung has multiple proposed cellular development pathways, and future studies that describe the molecular profiles of the lung cell types or lung cancer stem cells would further clarify the putative ancestral cells of the SCC subtypes (50).

The SCC subtypes may have applications in patient care and in cancer research. For instance, patients with the primitive subtype could be treated more aggressively because of the poor survival expectation of this subtype or could be given a more accurate prognosis than by using traditional prognostic factors alone. Basic cancer research could be conducted using the subtype model system partners described in this study. The SCC subtypes could be useful for therapy benefit studies and possibly serve as a foundation for clinical trial selection.

In conclusion, we identified four robust expression subtypes of lung SCC using a multicohort discovery and validation strategy. The subtypes are clinically and phenotypically different, suggesting different therapies.

Disclosure of Potential Conflicts of Interest

D.N. Hayes, C.M. Perou, and P.S. Bernard hold a provisional patent that is related to work described in this manuscript but there is no current financial interest. All other authors have no conflicts of interest.

Grant Support

NIH/National Cancer Institute grant F32CA142039 (M.D. Wilkerson), Lineberger Comprehensive Cancer Center Translational Small Grants Program (D.N. Hayes), NIH/National Center for Research Resources grant K12-RR023248 (D.N. Hayes), and Thomas G. Labrecque Foundation through Joan's Legacy Foundation (D.N. Hayes).

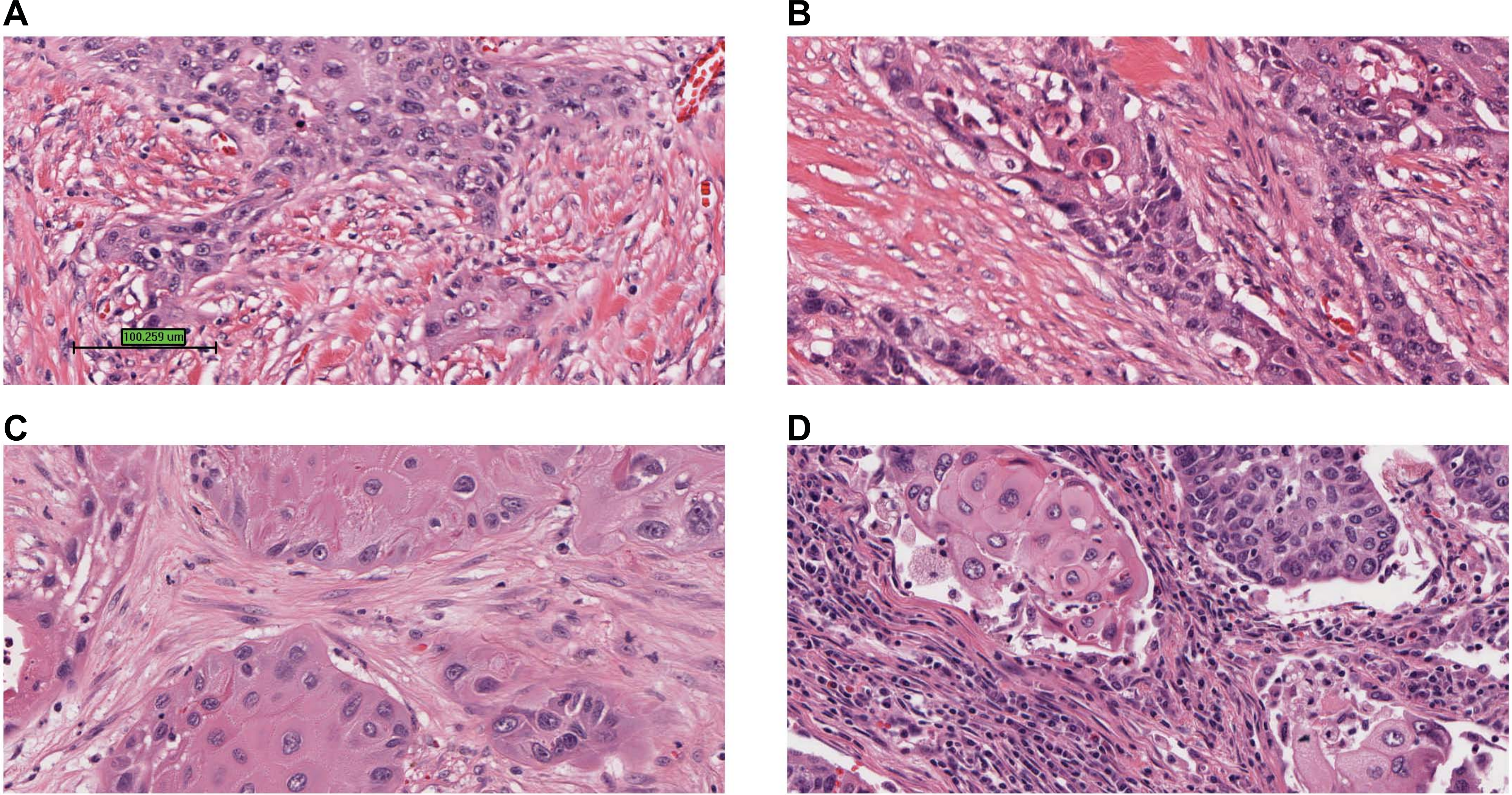
The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received 01/22/2010; revised 05/19/2010; accepted 06/27/2010; published OnlineFirst 07/19/2010.

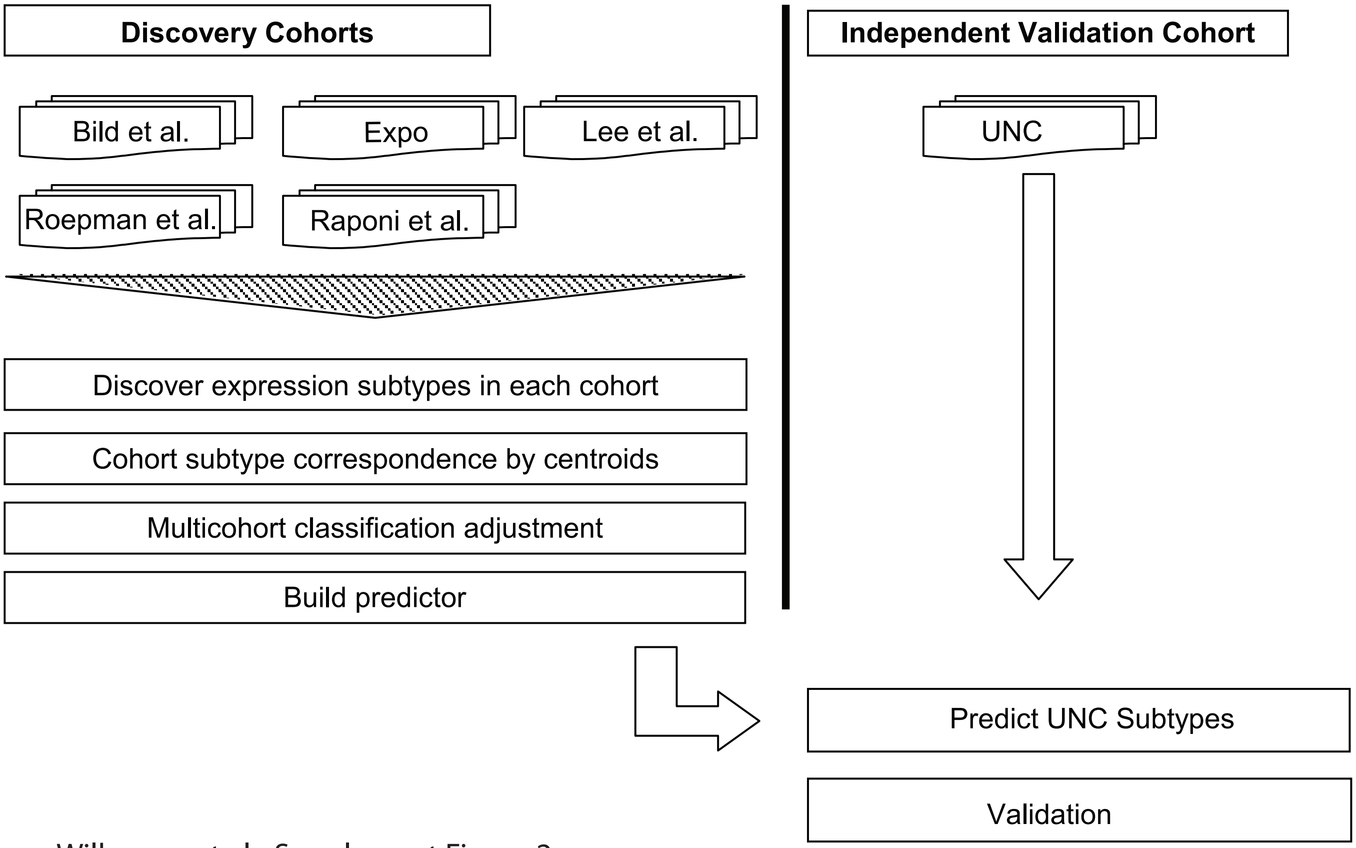
References

1. Parkin DM, Bray F, Ferlay J, Pisani P. Global cancer statistics, 2002. CA Cancer J Clin 2005;55:74–108.
2. Koyi H, Hillerdal G, Branden E. A prospective study of a total material of lung cancer from a county in Sweden 1997-1999: gender, symptoms, type, stage, and smoking habits. Lung Cancer 2002;36:9–14.
3. Visbal AL, Williams BA, Nichols FC III, et al. Gender differences in non-small-cell lung cancer survival: an analysis of 4,618 patients diagnosed between 1997 and 2002. Ann Thorac Surg 2004;78:209–15, discussion 15.
4. Travis WD, World Health Organization. Histological typing of lung and pleural tumours. Berlin: Springer; 1999.
5. Auerbach O, Garfinkel L, Parks VR. Histologic type of lung cancer in relation to smoking habits, year of diagnosis and sites of metastases. Chest 1975;67:382–7.
6. Jones DR, Detterbeck FC. Surgery for stage I non-small cell lung cancer. In: Detterbeck FC, Socinski MA, Rivera MP, Rosenman JG, editors. Diagnosis and treatment of lung cancer. 1st ed. Philadelphia: W.B. Saunders Company; 2001, p. 177–90.
7. Churg AM, Myers JL, Tazelaar HD, Wright JL. Thurlbeck's pathology of the lung. 3rd ed. New York: Thieme Medical Publishers, Inc.; 2005.
8. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531–7.
9. Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of human breast tumours. Nature 2000;406:747–52.
10. Hayes DN, Monti S, Parmigiani G, et al. Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts. J Clin Oncol 2006;24:5079–90.
11. Inamura K, Fujiwara T, Hoshida Y, et al. Two subclasses of lung squamous cell carcinoma with different gene expression profiles and prognosis identified by hierarchical clustering and non-negative matrix factorization. Oncogene 2005;24:7105–13.
12. Larsen JE, Pavey SJ, Passmore LH, et al. Expression profiling defines a recurrence signature in lung squamous cell carcinoma. Carcinogenesis 2007;28:760–6.
13. Raponi M, Zhang Y, Yu J, et al. Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. Cancer Res 2006;66:7466–72.
14. Hu Z, Troester M, Perou CM. High reproducibility using sodium hydroxide-stripped long oligonucleotide DNA microarrays. Biotechniques 2005;38:121–4.
15. Ritchie ME, Silver J, Oshlack A, et al. A comparison of background correction methods for two-colour microarrays. Bioinformatics 2007;23:2700–7.
16. Garrett-Mayer E, Parmigiani G, Zhong X, Cope L, Gabrielson E. Cross-study validation and combined analysis of gene expression microarray data. Biostatistics 2008;9:333–54.
17. Bild AH, Yao G, Chang JT, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 2006;439:353–7.
18. Lee ES, Son DS, Kim SH, et al. Prediction of recurrence-free survival in postoperative non-small cell lung cancer patients by using an integrated model of clinical information and gene expression. Clin Cancer Res 2008;14:7397–404.
19. Roepman P, Jassem J, Smit EF, et al. An immune response enriched 72-gene prognostic profile for early-stage non-small-cell lung cancer. Clin Cancer Res 2009;15:284–90.
20. Fischer AJ, Goss KL, Scheetz TE, Wohlford-Lenane CL, Snyder JM, McCray PB, Jr. Differential gene expression in human conducting airway surface epithelia and submucosal glands. Am J Respir Cell Mol Biol 2009;40:189–99.
21. Mariani TJ, Reed JJ, Shapiro SD. Expression profiling of the developing mouse lung: insights into the establishment of the extracellular matrix. Am J Respir Cell Mol Biol 2002;26:541–8.
22. Ross AJ, Dailey LA, Brighton LE, Devlin RB. Transcriptional profiling of mucociliary differentiation in human airway epithelial cells. Am J Respir Cell Mol Biol 2007;37:169–85.
23. Zhou BB, Peyton M, He B, et al. Targeting ADAM-mediated ligand cleavage to inhibit HER3 and EGFR pathways in non-small cell lung cancer. Cancer Cell 2006;10:39–50.
24. Gollub J, Sherlock G. Clustering microarray data. Methods Enzymol 2006;411:194–213.
25. Sorlie T, Tibshirani R, Parker J, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci U S A 2003;100:8418–23.
26. Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning 2003;52:91–118.

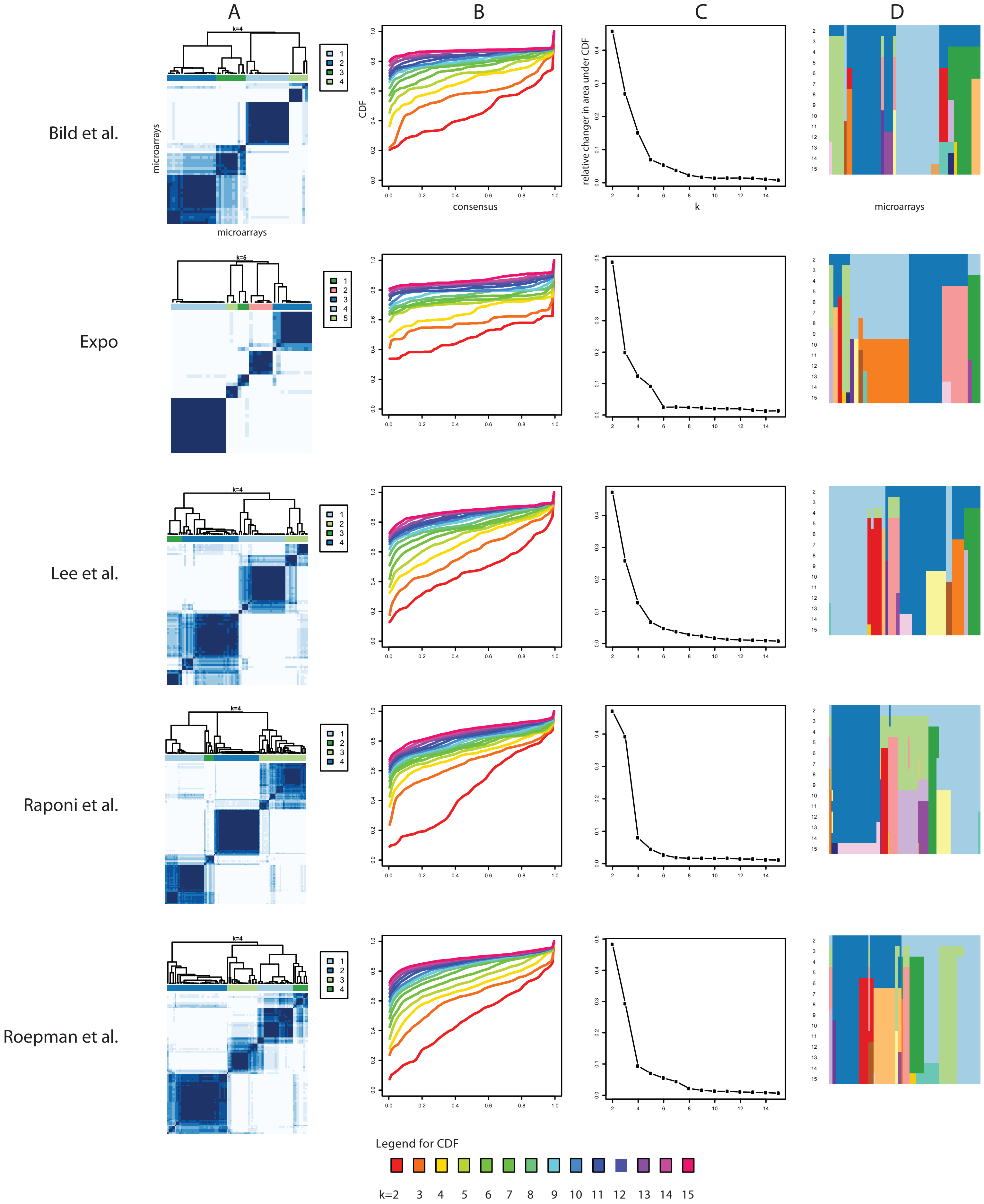
27. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics 2010;26:1572–3.
28. Hu Z, Fan C, Oh DS, et al. The molecular portraits of breast tumors are conserved across microarray platforms. BMC Genomics 2006;7:96.
29. Choi JK, Yu U, Kim S, Yoo OJ. Combining multiple microarray studies and modeling interstudy variation. Bioinformatics 2003;19 Suppl 1:i84–90.
30. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545–50.
31. Liu YH, Hayes DN, Nobel A, Marron JS. Statistical significance of clustering for high-dimension, low-sample size data. J Am Stat Assoc 2007:103.
32. Verhaak RGW, Hoadley KA, Purdom E, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 2010;17:98–110.
33. Cabanski CR, Qi Y, Yin X, et al. SWISS MADE: Standardized WithIn Class Sum of Squares to Evaluate Methodologies and Dataset Elements. PLoS One 5:e9905.
34. Whitfield ML, George LK, Grant GD, Perou CM. Common markers of proliferation. Nat Rev Cancer 2006;6:99–106.
35. Perou CM, Jeffrey SS, van de Rijn M, et al. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc Natl Acad Sci U S A 1999;96:9212–7.
36. Polager S, Ginsberg D. E2F—at the crossroads of life and death. Trends Cell Biol 2008;18:528–35.
37. Spira A, Beane J, Shah V, et al. Effects of cigarette smoke on the human airway epithelial cell transcriptome. Proc Natl Acad Sci U S A 2004;101:10143–8.
38. Jorgensen E, Stinson A, Shan L, Yang J, Gietl D, Albino AP. Cigarette smoke induces endoplasmic reticulum stress and the unfolded protein response in normal and malignant human lung cells. BMC Cancer 2008;8:229.
39. Chung CH, Parker JS, Karaca G, et al. Molecular classification of head and neck squamous cell carcinomas using patterns of gene expression. Cancer Cell 2004;5:489–500.
40. King KE, Weinberg WC. p63: defining roles in morphogenesis, homeostasis, and neoplasia of the epidermis. Mol Carcinog 2007;46:716–24.
41. Massion PP, Taflan PM, Jamshedur Rahman SM, et al. Significance of p63 amplification and overexpression in lung cancer development and prognosis. Cancer Res 2003;63:7113–21.
42. Gaillard D, Puchelle E. Differentiation and maturation of airway epithelial cells: role of extracellular matrix and growth factors. In: Gaultier C, Bourbon JR, Post M, editors. Lung development. Oxford: Oxford University Press; 1999, p. 46–76.
43. Garber ME, Troyanskaya OG, Schluens K, et al. Diversity of gene expression in adenocarcinoma of the lung. Proc Natl Acad Sci U S A 2001;98:13784–9.
44. Broome AM, Ryan D, Eckert RL. S100 protein subcellular localization during epidermal differentiation and psoriasis. J Histochem Cytochem 2003;51:675–85.
45. Smith SL, Gugger M, Hoban P, et al. S100A2 is strongly expressed in airway basal cells, preneoplastic bronchial lesions and primary non-small cell lung carcinomas. Br J Cancer 2004; 91:1515–24.
46. Chu PG, Weiss LM. Keratin expression in human tissues and neoplasms. Histopathology 2002;40:403–39.
47. Ozasa H, Oguri T, Uemura T, et al. Significance of thymidylate synthase for resistance to pemetrexed in lung cancer. Cancer Sci 101:161–6.
48. Wuenschell CW, Sunday ME, Singh G, Minoo P, Slavkin HC, Warburton D. Embryonic mouse lung epithelial progenitor cells co-express immunohistochemical markers of diverse mature cell lineages. J Histochem Cytochem 1996;44:113–23.
49. Prat A, Perou CM. Mammary development meets cancer genomics. Nat Med 2009;15:842–4.
50. Snyder JC, Teisanu RM, Stripp BR. Endogenous lung stem cells and contribution to disease. J Pathol 2009;217:254–64.



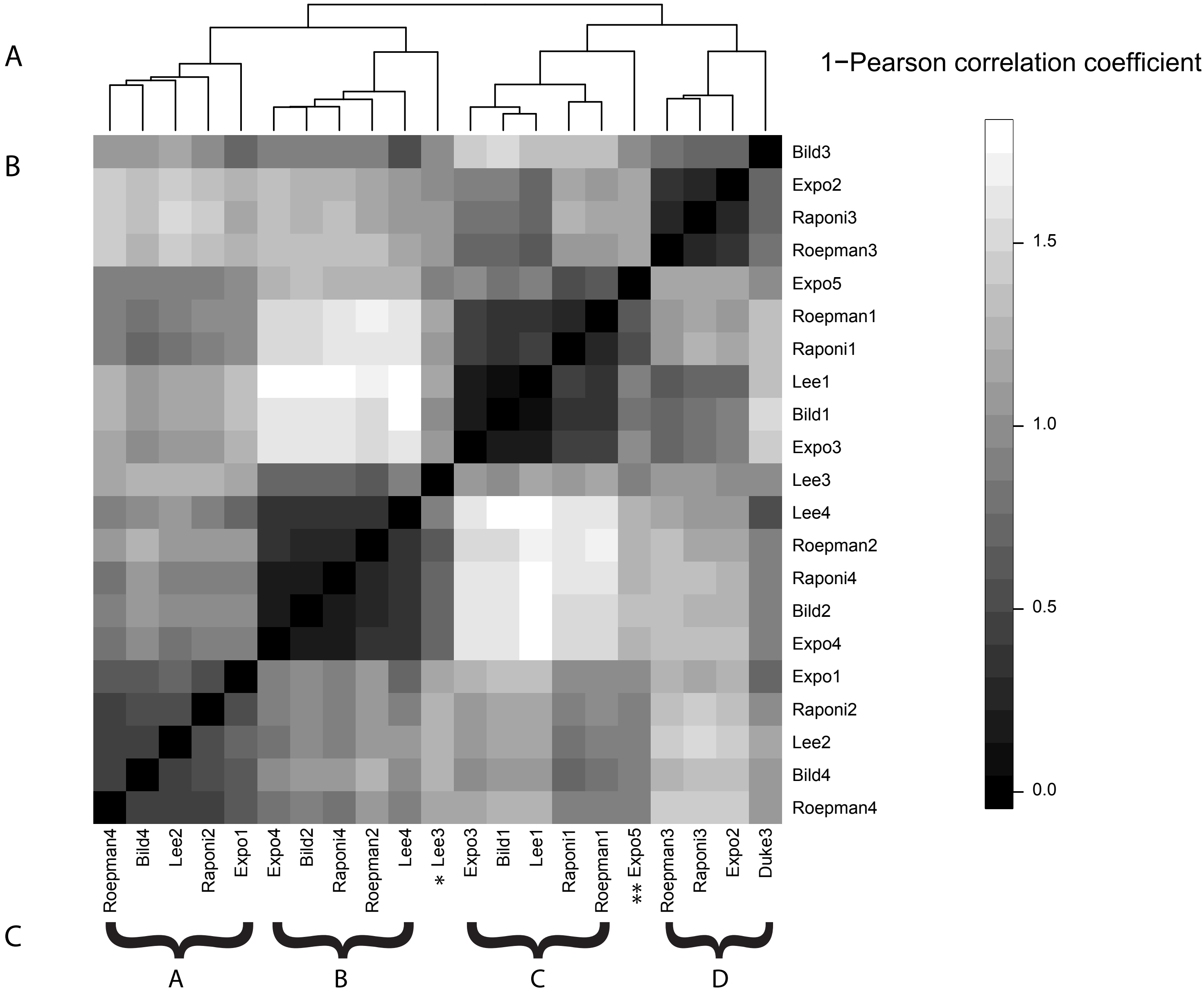
Wilkerson et al - Supplement Figure 1.


Discovery Cohorts
Independent Validation Cohort
Bild et al.
Expo
Lee et al.
UNC
Roepman et al.
Raponi et al.
Discover expression subtypes in each cohort
Cohort subtype correspondence by centroids
Multicohort classification adjustment
Build predictor
Predict UNC Subtypes
Validation

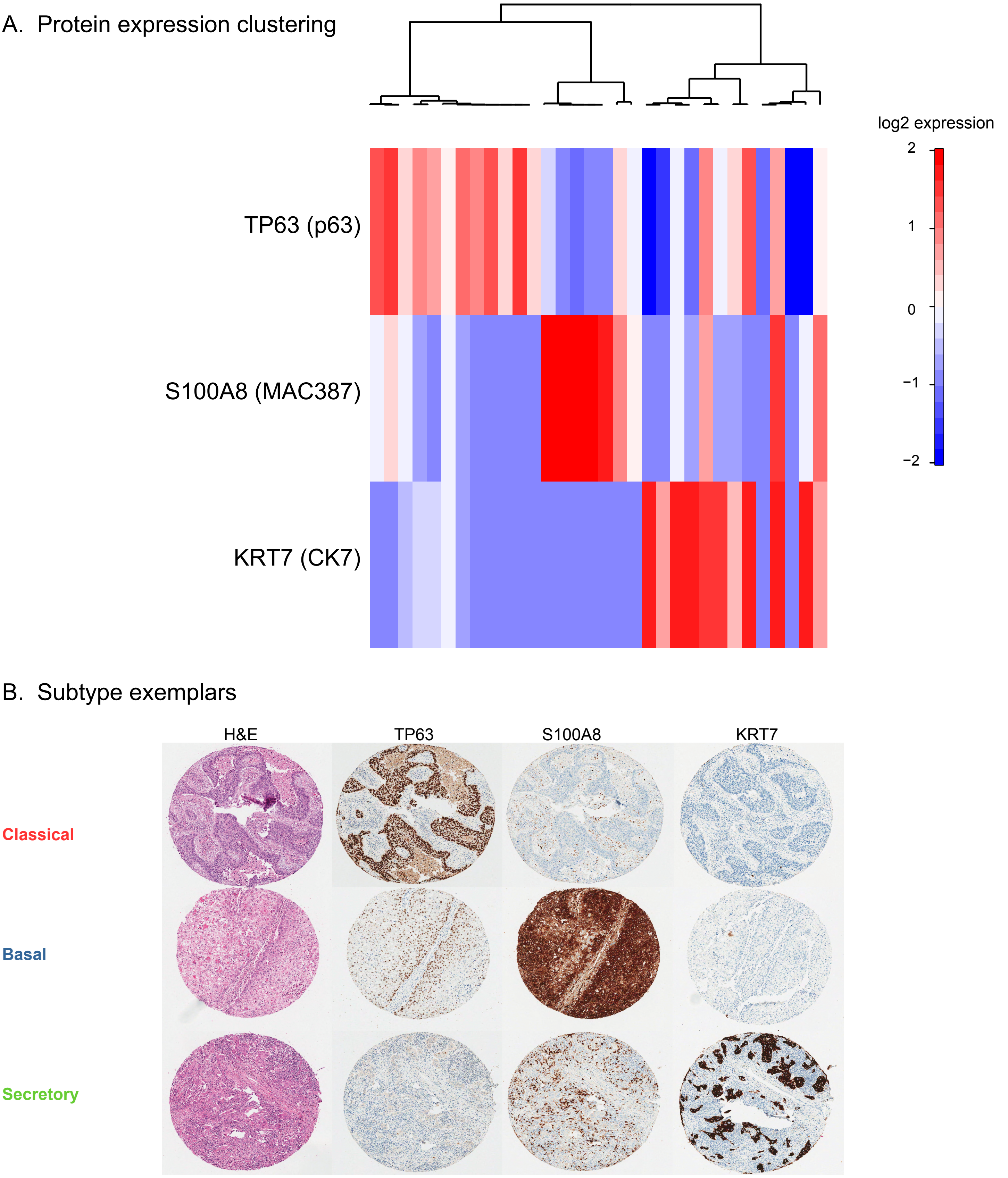
Wilkerson et al - Supplement Figure 2.



Wilkerson et al - Supplement Figure 3.

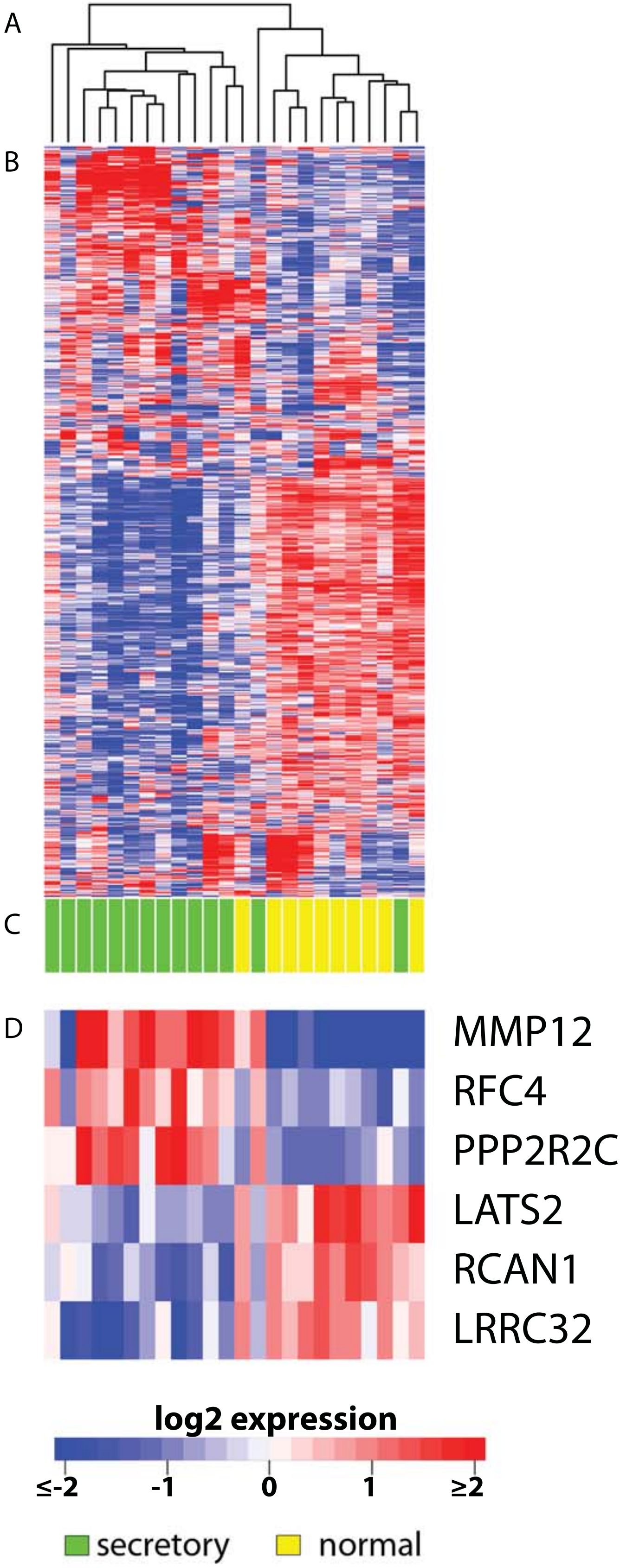


Wilkerson et al - Supplement Figure 4.

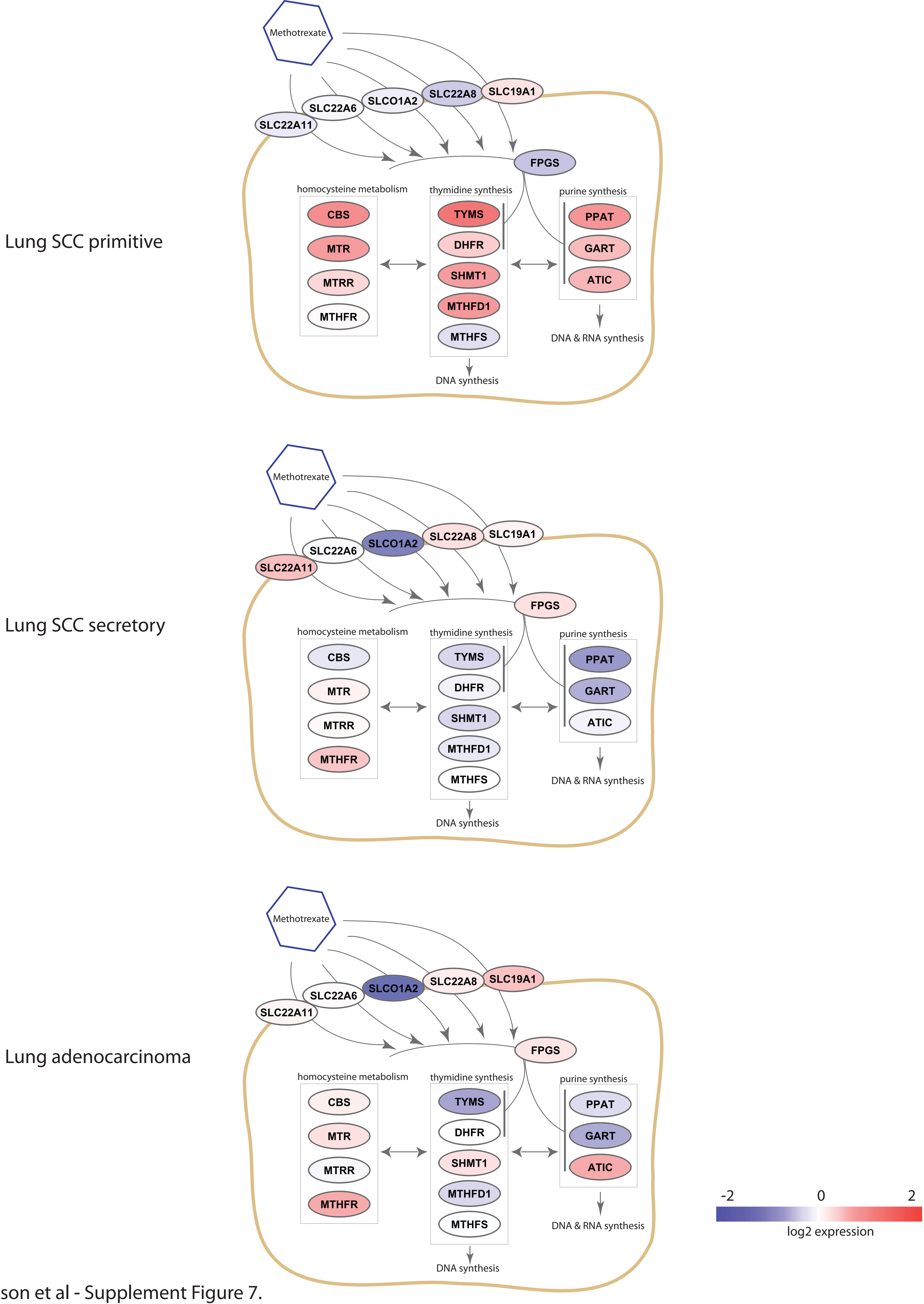

A. Protein expression clustering

B. Subtype exemplars

Wilkerson et al - Supplement Figure 5.



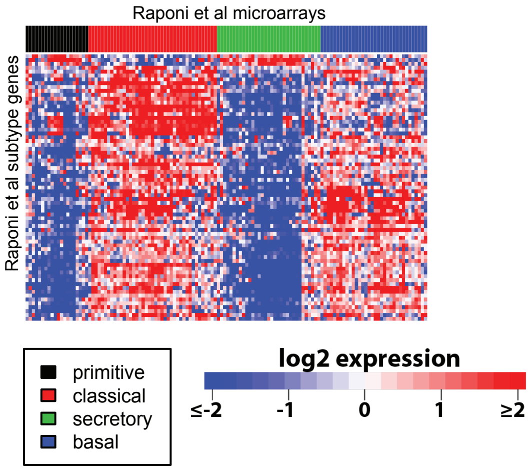
Wilkerson et al - Supplement Figure 6.


Lung SCC primitive

Lung SCC secretory

Lung adenocarcinoma

Wilkerson et al - Supplement Figure 7.



Wilkerson et al - Supplement Figure 8.

Supplementary material for:

Wilkerson, M.D. et al. (2010) Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically-important and correspond to different normal cell types.

Acquisition, quality control, and processing of published lung SCC microarrays

Clinical and microarray data were downloaded from websites for the Bild et al, Expo, Raponi et al, and Roepman et al discovery cohorts (Supplemental Table 1). Arrays were unbiasedly reviewed for possible technical artifacts suggestive of probe hybridization irregularities using spatial intensity, overall intensity, relative log expression, normalized unscaled errors, mRNA degradation and overall intensity plots. Arrays showing evidence of possible technical artifacts were removed from further analysis. Discovery cohorts were reduced to those with an SCC diagnosis. The following SCC microarrays were removed: Bild et al - 0176_6612_h133+_97-403.cel; Expo - GSM231874; Raponi et al – GSM102217, GSM102215. The removed Raponi et al microarrays were noted in the original publication as having reduced quality. All microarray platform probes were mapped to a common gene database to create gene expression values. A database of curated mRNA transcripts corresponding to human genome build 36.1 and GenBank release 161 was downloaded[1] (1). Separately for the Affymetrix U133 Plus 2.0, Affymetrix U133A and UNC custom Agilent 44,000 platforms, probes were aligned to transcripts by BLAT (2) and probes with completely identical, same strand, no-gap alignments to exactly one gene were retained. Raponi et al array transcripts were aligned to this database and transcripts with 90% identical, no-gap alignments to exactly one gene were retained. Roepman et al Unigene identifiers were mapped to gene symbols. For cohorts with Affymetrix CEL files, expression values were calculated using the Robust Multiarray Average (3) and the custom mapping. Otherwise, probes or transcripts matching the same gene were averaged. Raponi et al expression values were log2 transformed to be on the same scale as the other cohorts. Final platform and cohort gene counts are listed in Supplementary Table 1. Cohort clinical variable levels were mapped to common scales where needed including grades moderate-poor and moderate-well mapped to moderate and age range mapped to the range mean.

Ross et al microarrays were processed into gene expression values by Robust Multi-Array Average (3) and the custom gene mapping. Ross et al microarrays were reduced to the common time points among the three patients. Patient median gene expression was calculated for each time point. Mariani et al mouse microarrays were processed by Robust Multi-Array Average (3). Human-mouse homologous genes were downloaded from NBCI Homologene[2]. Mariani et al genes were mapped to human homologs and genes not in a one-to-one relationship were removed. All SCC cell lines from Zhou et al microarrays were processed into gene expression values by Robust Multi-Array Average (3) and the custom gene mapping. Ross et al, Mariani et al, and Zhou et al data were gene median centered. Final gene counts are in Supplementary Table 1.

Unsupervised subtype discovery and multi-cohort classification adjustment.

[1] ftp://ftp1.nci.nih.gov/tcga/other/integration/db/SpliceMiner_9606TranscriptDB_36.1.zip

[2] ftp://ftp.ncbi.nih.gov/pub/HomoloGene/; version Feb. 14, 2008

Consensus Clustering provides quantitative stability evidence for judging the number of clusters in a microarray dataset (4). This stability evidence, termed consensus, is the proportion that two microarrays are clustered together over a large number of microarray subsamplings. All discovery cohorts' consensus empirical cumulative distributions have modes near 0 and 1 (Supplemental Fig 3B) indicating that tumors have high consensus to some tumors and low consensus to others which is evidence for clusters (4). Consensus proportional increases approached a minimum at four clusters in all cohorts, which shows that additional clusters are similar to random divisions (Supplemental Fig. 3B, C). Consensus matrices demonstrate high intra-cluster consensus and low inter-cluster consensus at four clusters, confirming four as a stable cluster number (Supplemental Fig. 3A). All cohorts' cluster tracking plots demonstrated that each of four clusters comprised > 10% of samples in a cohort and that additional clusters were small (Supplemental Fig. 3D). The Expo dataset had an equivalently sized 5th cluster. A likely cause for this additional Expo cohort cluster is that this cohort is the smallest and the paucity of samples complicates detection of exactly 4 clusters as in the other cohorts. By sum of this evidence, four clusters were selected as a common, empirically-supported number of expression clusters in all discovery cohorts.

In order to derive the optimal sample classification given all of the data rather than data from its source cohort, a multi-cohort classification step was completed. Multi-cohort centroids were built by taking the median of each centroid group (A, B, C, D in Supplemental Fig 4). Then, all arrays were classified by taking the maximum Pearson correlation to these multi-cohort centroids. After this adjustment, group D was found in the Lee et al dataset. We note that group D was also found in Lee et al at a higher consensus cluster count (data not shown). An average of 15% of a cohort's arrays changed classification; thus, a minority of arrays had their classification modified.

The Raponi et al cohort contained one patient assayed by two microarrays. Both arrays were the same subtype and we retained one patient record for clinical analysis.

Supplement Table 1: Data source, probe annotation and array counts.

	Lung SCC Patient Cohorts						Model Datasets		
	UNC	**Bild et al.** [1]	**Expo** [2]	**Lee et al.** [3]	**Raponi et al.** [4]	**Roepman et al.** [5]	**Ross et al** [6]	**Mariani et al** [7]	**Zhou et al** [8]
Institution	University of North Carolina	Duke University	International Genomics Consortium	Sungkyunkwan University	University of Michigan	European Microarray Consortium			
Microarray platform	Agilent 44K custom	Affymetrix U133 Plus 2.0	Affymetrix U133 Plus 2.0	Affymetrix U133 Plus 2.0	Affymetrix U133A	Agilent 44K whole genome	Affymetrix U133 Plus 2.0	Affymetrix Mu11K subA and subB	Affymetrix U133A
Expression level	probe	probe	probe	probe	gene	probe	probe	probe	probe
Expression format	sample / common reference	Affymetrix CEL files	Affymetrix CEL files	Affymetrix CEL files	MAS5	Log2 ratio (sample/common reference)	Affymetrix CEL files	Affymetrix CEL files	Affymetrix CEL files
Published annotation	probe sequences	probe sequences	probe sequences	probe sequences	Affymetrix Transcript	Unigene and other	probe sequences	probe sequences	probe sequences
Lung squamous cell carcinoma array count	56	53	37	75	130	92	-	-	4
Array count with acceptable quality control	-	52	36	75	128	92	30	11	4
Probes/Transcripts on array	39,980	604,258	604,258	604,258	22,283	44,290	604,258	8,828	247,965
Probes on array mapping to exactly one gene	31,035	318,205	318,205	318,205	-	-	318,205	-	195,448
Transcripts mapping to exactly one gene	-	-	-	-	17,320	29,734	-	6,286 [9]	
Final Gene count	17,109	17,537	17,537	17,537	11,865	15,263	17,537	6,286	12,301
Genes in common across cohorts	9,515								
Genes meeting reliability condition across cohorts	9,229								

1. http://data.genome.duke.edu/oncogene.php and (5)
2. ftp://ftp.ncbi.nih.gov/pub/geo/DATA/supplementary/series/GSE2109/GSE2109%5FRAW%2Etar
3. (6)
4. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4573 and (7)
5. http://research.agendia.com/ and (8)
6. (9)
7. (10)
8. (11)
9. Human ortholog gene counts.

	Overall survival Hazard ratio (95% CI)		Relapse-free survival Hazard ratio (95% CI)	
univariate				
Primitive subtype vs. others	1.83	(1.18-2.86)*	2.40	(1.49-3.86)*
Stage	1.24	(1.10-1.40)*	1.15	(0.98-1.36)
Grade	1.12	(0.65-1.94)	1.73	(0.97-3.10)
Age ≥ 70	0.93	(0.64-1.37)	1.03	(0.63-1.69)
multivariate				
Primitive subtype vs others	1.95	(1.11-3.43)*	2.03	(1.08-3.80)*
Stage	1.17	(0.99-1.37)	1.13	(0.92-1.38)
Grade	0.98	(0.56-1.72)	1.43	(0.77-2.65)
Age ≥ 70	0.84	(0.51-1.41)	1.14	(0.60-2.17)

Supplement Table 2: Cox proportional hazards models. Cox proportional hazards models used all available data, have tumor stage coded as a number 1-7 for stage IA through IV, grade coded as poorly differentiated or other, and age coded as greater than or equal to the median patient age, 70 (* $P < 0.05$).

Figure Legends.

Supplement Figure 1: Subtype exemplar H&E images.
Pathologist-reviewed exemplars of each subtype (A – primitive, B- classical, C – secretory, D – basal) are displayed. The scale in A also applies to B-D.

Supplement Figure 2: SCC subtype discovery and validation procedure.

Supplement Figure 3: Consensus clustering of discovery cohorts.
Consensus clustering results from the discovery cohorts are shown as rows. Consensus matrices are symmetrical and represent consensus values at a particular cluster count (*k*) between two microarrays (A). Consensus is the proportion that two microarrays occur in the same cluster out of number of subsample iterations. Consensus is shown according to the color range of dark blue for a consensus value of 1 and white for a consensus value of 0. The clusters are indicated by the colored rectangles atop the matrix according to the color legend within each cohort (A). Empirical cumulative consensus distributions are shown for different *k* (B). Consensus proportional increase plots show the change in area under the curve in (B) in comparing *k* relative to *k*-1 (C). Item tracking plots show the cluster assignment of microarrays in columns over different *k* clusterings, colors indicate the same cluster (D). The consensus matrices' clusters colors correspond to the cluster tracking plot colors. For further details, refer to the Consensus Clustering publication (4).

Supplement Figure 4: Correlation matrix and clustering of unsupervised centroids from discovery cohorts.
Cells are labeled by discovery cohort and centroid where the centroid number is taken from unsupervised clustering (Supplement Fig. 3). Cells in the matrix represent the 1 – Pearson correlation coefficient between two discovery cohort centroids by a degree of shading according to the scale above (B). For example, Roepman4 and Bild2 have highly similar expression profiles, have a large Pearson correlation coefficient, a small 1 – Pearson correlation coefficent value and is shaded darkly. The matrix is ordered by columns and rows by the dendrogram at the top of the matrix (A). The dendrogram is the result of an agglomerative, average linkage, hierarchical clustering using the correlation matrix. Four centroid groups are marked (C). All cohorts have one member in each centroid group with one exception: Lee et al. does not have a centroid in group D. Lee et al. has an extra centroid, Lee3, clustered with group B that is less similar to the group than Lee4, and so Lee3 excluded from this centroid group (*). Expo5 represents a small cluster of 3 microarrays (**). Because Expo5 is clustered with group C and Expo3 is more similar to the group, Expo5 is not included in centroid group D.

Supplement Figure 5: Protein expression by immunohistochemistry. Protein expression was evaluated by pathologist review of immunohistochemical staining intensity of tumor cells via a tissue microarray. Scores are the proportion of tumor cells (0-100) multiplied by their immunohistochemical stain intensity (0-3). Scores were standardized prior to agglomerative average-linkage hierarchical clustering and are displayed as a heatmap in which columns are tumor samples (A), and rows are genes targeted by antibodies shown in parentheses. One exemplar per subtype is displayed in rows and immunohistochemical stains in columns (B).

Supplement Figure 6: Unsupervised clustering of squamous secretory subtype and normal microarrays. The top 1,000 variable genes, measured by median absolute deviation, were used for unsupervised clustering (A) and heatmap display (B). The clustering was agglomerative, average linkage, hierarchical clustering using 1-Pearson correlation coefficient as distance. Microarrays were gene-median centered prior to clustering. Secretory subtype and normal microarrays are marked by the colored rectangles (C) according to the legend. Representative genes are shown (D).

Supplement Figure 7: Comparison of methotrexate (antifolate) drug metabolism pathway among lung squamous subtypes. Pathway is derived from PharmGKB Methotrexate drug metabolism pathway[3]. SCC subtype centroid gene expression is represented by the color scale. An adenocarcinoma centroid is presented for comparison. The adenocarcinoma centroid has the squamous gene medians subtracted, so that it is on the same scale as the squamous subtype centroids. Expression data is from UNC cohort and unpublished local adenocarcinoma samples.

Supplement Figure 8: Comparison to previously published subtypes. A heatmap shows expression of the Raponi et al subtype genes, as rows, for the Raponi et al microarrays, as columns. Microarrays are grouped and labeled by the subtypes defined in this study, indicated by the colored rectangles.

[3] http://www.pharmgkb.org/do/serve?objId=PA2039

References

1. Kahn AB, Ryan MC, Liu H, Zeeberg BR, Jamison DC, Weinstein JN. SpliceMiner: a high-throughput database implementation of the NCBI Evidence Viewer for microarray splice variant analysis. BMC Bioinformatics 2007;8: 75.
2. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res 2002;12: 656-64.
3. Irizarry RA, Hobbs B, Collin F*, et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003;4: 249-64.
4. Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning 2003;52: 91-118.
5. Bild AH, Yao G, Chang JT*, et al.* Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 2006;439: 353-7.
6. Lee ES, Son DS, Kim SH*, et al.* Prediction of recurrence-free survival in postoperative non-small cell lung cancer patients by using an integrated model of clinical information and gene expression. Clin Cancer Res 2008;14: 7397-404.
7. Raponi M, Zhang Y, Yu J*, et al.* Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. Cancer Res 2006;66: 7466-72.
8. Roepman P, Jassem J, Smit EF*, et al.* An immune response enriched 72-gene prognostic profile for early-stage non-small-cell lung cancer. Clin Cancer Res 2009;15: 284-90.
9. Ross AJ, Dailey LA, Brighton LE, Devlin RB. Transcriptional profiling of mucociliary differentiation in human airway epithelial cells. Am J Respir Cell Mol Biol 2007;37: 169-85.
10. Mariani TJ, Reed JJ, Shapiro SD. Expression profiling of the developing mouse lung: insights into the establishment of the extracellular matrix. Am J Respir Cell Mol Biol 2002;26: 541-8.
11. Zhou BB, Peyton M, He B*, et al.* Targeting ADAM-mediated ligand cleavage to inhibit HER3 and EGFR pathways in non-small cell lung cancer. Cancer Cell 2006;10: 39-50.