VOLUME 24 · NUMBER 31 · NOVEMBER 1 2006

JOURNAL OF CLINICAL ONCOLOGY

REVIEW ARTICLE

Gene Expression Profiling Reveals Reproducible Human Lung Adenocarcinoma Subtypes in Multiple Independent Patient Cohorts

D. Neil Hayes, Stefano Monti, Giovanni Parmigiani, C. Blake Gilks, Katsuhiko Naoki, Arindam Bhattacharjee, Mark A. Socinski, Charles Perou, and Matthew Meyerson

From the Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC; Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge; Departments of Medical Oncology and Pathology, Dana-Farber Cancer Institute, Harvard Medical School, Boston; Agilent Technologies, Andover, MA; Department of Biostatistics, The Johns Hopkins University School of Medicine, Baltimore, MD; Department of Pathology and Laboratory Medicine, Vancouver General Hospital and University of British Columbia, Vancouver, British Columbia, Canada; and Yokohama Municipal Hospital, Yokohama, Japan.

Submitted December 14, 2005; accepted August 18, 2006.

Authors' disclosures of potential conflicts of interest and author contributions are found at the end of this article.

Address reprint requests to D. Neil Hayes, MD, MPH, Assistant Professor of Medicine, University of North Carolina, Lineberger Comprehensive Cancer Center, CB #7295, Chapel Hill, NC 27599-7295; e-mail: hayes@med.unc.edu.

0732-183X/06/2431-5079/$20.00

DOI: 10.1200/JCO.2005.05.1748

ABSTRACT

Purpose
Published reports suggest that DNA microarrays identify clinically meaningful subtypes of lung adenocarcinomas not recognizable by other routine tests. This report is an investigation of the reproducibility of the reported tumor subtypes.

Methods
Three independent cohorts of patients with lung cancer were evaluated using a variety of DNA microarray assays. Using the integrative correlations method, a subset of genes was selected, the reliability of which was acceptable across the different DNA microarray platforms. Tumor subtypes were selected using consensus clustering and genes distinguishing subtypes were identified using the weighted difference statistic. Gene lists were compared across cohorts using centroids and gene set enrichment analysis.

Results
Cohorts of 31, 72, and 128 adenocarcinomas were generated for a total of 231 microarrays, each with 2,553 reliable genes. Three adenocarcinoma subtypes were identified in each cohort. These were named bronchioid, squamoid, and magnoid according to their respective correlations with gene expression patterns from histologically defined bronchioalveolar carcinoma, squamous cell carcinoma, and large-cell carcinoma. Tumor subtypes were distinguishable by many hundreds of genes, and lists generated in one cohort were predictive of tumor subtypes in the two other cohorts. Tumor subtypes correlated with clinically relevant covariates, including stage-specific survival and metastatic pattern. Most notably, bronchioid tumors were correlated with improved survival in early-stage disease, whereas squamoid tumors were associated with better survival in advanced disease.

Conclusion
DNA microarray analysis of lung adenocarcinomas identified reproducible tumor subtypes which differ significantly in clinically important behaviors such as stage-specific survival.

J Clin Oncol 24:5079-5090. © 2006 by American Society of Clinical Oncology

INTRODUCTION

Lung cancer is the leading cause of cancer death worldwide.[1] Although a useful term for epidemiologic purposes, lung cancer does not refer to a specific disease, but rather represents a heterogeneous collection of tumors of the lung, bronchus, and pleura.[2] In clinical practice, however, most patients are designated to either the specific histologic diagnosis of small-cell lung carcinoma (SCLC) or the diagnosis of exclusion, non–small-cell lung carcinoma (NSCLC). The distinction, although crude, is useful due to striking differences in disease behavior and response to treatment.[3,4] The subclassification of the nonspecific diagnosis NSCLC for 80% of lung cancer patients is essential when viewed in light of the push toward targeted cancer therapy. The major histologic subtypes of NSCLC include adenocarcinomas (the most common form of lung cancer), squamous cell lung carcinomas (SQ), and large-cell lung carcinomas (LCLC).[2]

Within the category of adenocarcinoma of the lung, expert panels have recognized a number of subtypes and histologic variants. Most notably, the WHO's most recent edition of the *Histologic Typing of Lung and Pleural Tumors* describes no fewer than 13 diagnostic classifications.[2] With the exception of the tumor subtypes bronchioalveolar carcinoma (BAC) and adenocarcinoma with BAC features and their associated mutations of the epidermal growth

factor receptor (*EGFR*) gene, histologic subtypes and molecular markers have had little impact on clinical practice for NSCLC, with treatment based primarily on clinical stage.[5-7] Histologic subtypes have demonstrated interobserver variability too high for integration into routine practice, although the new WHO classification scheme offers promise for more reproducible diagnosis.[8-11]

In response to the need to develop useful tumor subtypes, researchers have turned to high-throughput screening assays such as DNA microarrays. These tools allow investigators to measure thousands of potential biomarkers for a given patient or cohort of patients in a single assay.[12] Two types of screening methods exist: either an exploration of genes associated with a specific outcome (ie, survival), or a global survey to elicit dominant patterns of gene expression without regard to a specific outcome, called clustering. When tumors cluster, they share a common biologic base, such as a genetic mutation. In a dramatic example, dominant gene expression patterns have demonstrated breast cancer subtypes reproducibly that mirror clinically important tumors genotypes and phenotypes, including estrogen receptor status, *BRCA* status, Her2/*neu* expression, and survival.[13-15]

In the field of lung cancer, microarray analysis by independent investigators has demonstrated a wide variety of potentially clinically important uses, including the ability to distinguish morphologic variants reliably and predict prognosis.[16-44] However, progress in the field has been slow in terms of clarifying the heterogeneity of tumor behavior, such as has been done extensively in breast cancer.[45] The state of gene expression profiling in lung cancer is probably best summarized by Takeuchi et al[16]: "To date, various groups including our own have reported that expression profiling can recapitulate morphologic classification of NSCLCs, and some studies also showed that adenocarcinomas can be subclassified additionally. However, these previously reported subclassifications vary considerably from study to study, making it difficult to reconcile their findings or reach any definite conclusions."

The challenges in reconciling results across gene expression studies are formidable. There is no consensus on the number of subgroups, with investigators reporting between two and more than six subtypes of adenocarcinomas. Furthermore, in the few cases where genes defining subgroups have been reported, the concordance across studies approaches 0%. Although clinical, molecular, and morphologic characteristics have been reported to vary by subtype, no association has emerged that would allow confident identification of adenocarcinoma subtypes in new data or mapping of subtypes across different studies. In summary, although lung cancer subtypes seem to exist, there is little consensus on their number and nature, and how they might be reidentified in a prospective manner. In our current work, we do not propose to repeat individual clustering analyses reported previously, but rather to build on the collective body of work. We hypothesize that through the use of a standardized and systematic method, clearly identifiable subtypes of lung adenocarcinoma can be demonstrated in multiple independent clinical patient cohorts. We propose that the reproducibility constitutes a validation of these tumor subtypes and we provide the means for future investigators to identify these clinically relevant tumor subtypes in a platform-independent manner.

METHODS

Tumor Samples

Multiple lung carcinoma microarray datasets have reported tumor subtypes, but direct comparisons of gene expression profiling studies have not been reported.[20-22] Therefore, we examined the three largest of these studies from the investigators at the University of Michigan (Michigan; Ann Arbor, MI), Stanford University (Stanford; Stanford, CA), and the Dana-Farber Cancer Institute (Dana-Farber; Boston, MA) reporting subtypes of lung adenocarcinoma as defined by expression profiling, and performed a coordinated analysis. Although the tumor of primary interest in the analysis was adenocarcinoma, other tumor and normal tissues were represented in these arrays, including normal lung (NL), SQ, SCLC, LCLC. Adenocarcinomas with the following characteristics were excluded because they were not universally represented across datasets: lymph node metastases of primary tumors, intrapulmonary metastases, distant metastases, and suspected colon metastases. Tumor morphologic type, including BAC status, was determined at the sponsoring institution for each dataset. It is not possible in these data to distinguish samples with pure BAC from those that might better be described as adenocarcinoma with BAC features. Construction of the histologically comparable cohort as well as links to all phenotype data on all samples is documented in the Supplementary Data (available online at http://www.jco.org).

Microarray Data Analysis

The following microarray platforms were used: Michigan, Affymetrix hu6800 GeneChip (Santa Clara, CA); Dana-Farber, 95av2 GeneChip (Affymetrix GeneChip); and Stanford, printed cDNA array using the IMAGE clone set (printed at Standford University, Stanford, CA; IMAGE clone set, Livermore, CA). All arrays were screened for quality by standard methods and experiments not meeting objectively defined quality thresholds were excluded. Quality screening is described in detail, including accounting of all excluded samples, in the Supplementary Data. Gene expression was computed for the oligonucleotide arrays using the robust multichip averaging method, whereas the Stanford Microarray Database Server (SMD) provided expression values for the cDNA arrays.[46,47] Arrays from the SMD server were processed as in the original report of the data.[21] To normalize gene expression for cross-platform comparisons, all genes were mean-centered within each sample set.[14,48] Unigene cluster identifiers were used to match the probes and probe sets to their representative genes.[49] Genes present on all three array formats were evaluated for cross-platform reliability using the unbiased method of integrative correlations (ICs).[18,50] Genes with IC coefficients exceeding 2 standard deviations above that expected by chance were considered reliable and used for additional analysis. Links to both raw and processed datasets are available in the Supplementary Data.

Robust clusters or tumor subtypes were selected in a standardized manner independently for each dataset (Fig 1). We used the consensus clustering method, which incorporates average linkage agglomerative hierarchical clustering using a widely accepted distance measure, 1 – (Pearson's correlation coefficient).[51] Confirmation of the optimal clustering assignments was by the independent clustering method, nonnegative matrix factorization, proposed by Brunet et al.[52] Having assigned all adenocarcinoma samples to their respective consensus clusters, we characterized the groups using the centroid method developed by Sørlie et al (see Appendix; online only).[14] Centroids were prepared for the following groups of samples: each adenocarcinoma consensus cluster subtype within the three cohorts, NL, SCLC, SQ, LCLC, and BAC. When a histologic group was present in multiple sample sets (such as NL), a separate centroid was prepared for each dataset in which it appeared. The NL, SQ, SCLC, LCLC, and BAC centroids were used as common references across platforms. Hierarchical agglomerative clustering and probabilistic clustering were used to detect correlations between centroids using the same distance measure as above.

Subtype Gene Lists

Lists of genes most closely associated with the adenocarcinoma clusters were generated using the statistical analysis of microarrays method (SAM; see Appendix).[53] SAM parameters were set to select genes associated with the subclasses in the one versus all, and all pair-wise comparisons, with a fixed false discovery rate (FDR) of 0.1%. If no genes were selected at an FDR of 0.1%, the criterion was relaxed iteratively until at least 10 genes were selected, with the algorithm recording the FDR at which the target was finally reached. In cases requiring relaxing the FDR, the degree of adjustment was

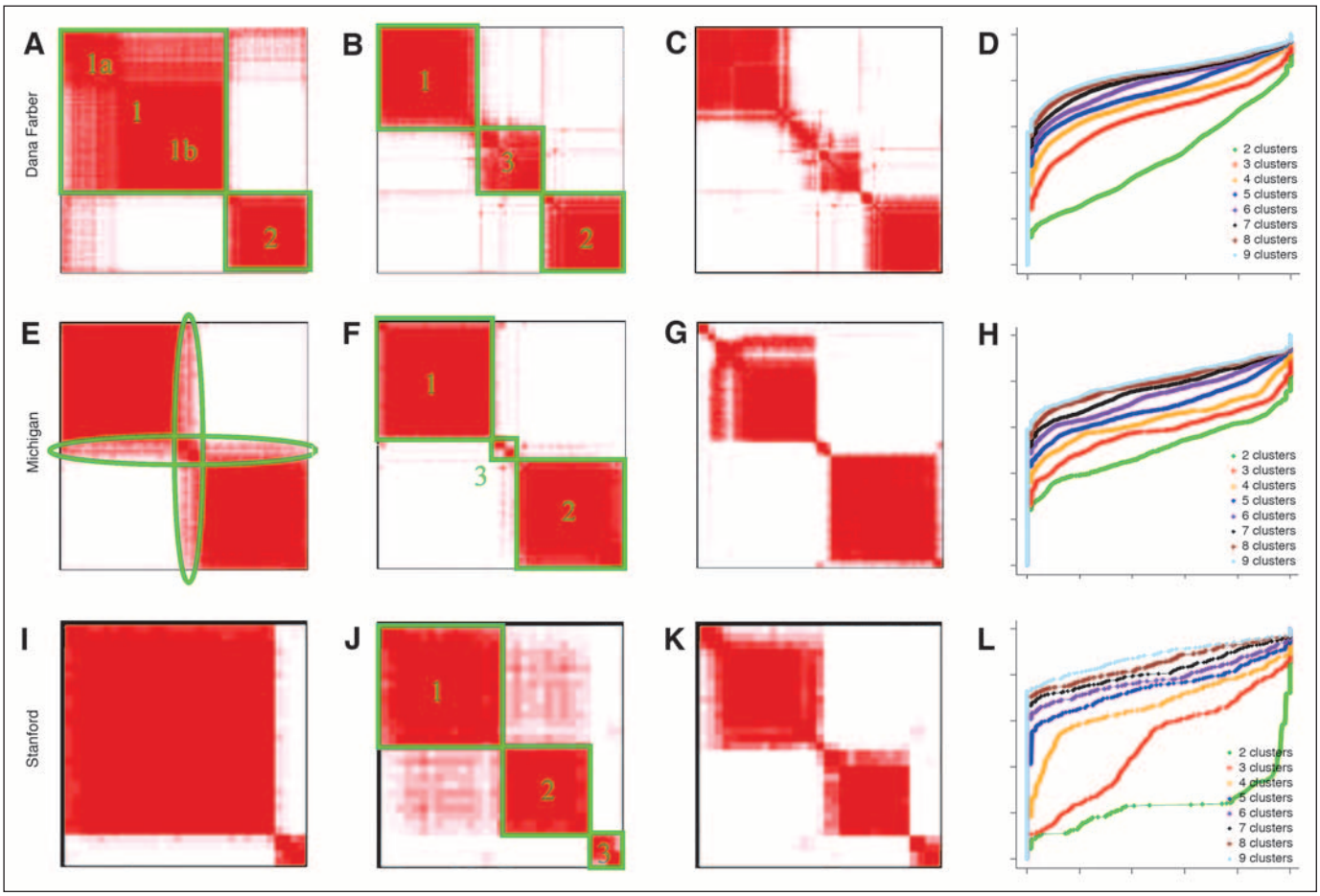

Fig 1. Consensus matrix by data set and cluster number. Rows correspond to each independently analyzed dataset. The first three columns from the left are the consensus matrices (CMs) for increasing numbers of **K** = 2 to 4 clusters. Each CM represents the frequency with which samples occur in the same grouping by hierarchical clustering pruned to $\mathbf{K_i}$ clusters. See Appendix for detailed discussion of consensus clustering.

suggested automatically by the delta statistic of the SAM algorithm. The result of this FDR adjustment strategy was that in cases where only a few genes are selected, the FDR was generally low. In cases of sparse data, however, the outcome occasionally was the selection of a large number of genes with a high FDR. Gene lists generated in this way were compared across datasets both in terms of their expected concordance and by the nonparametric methodology known as Gene Set Enrichment Analysis (GSEA; see Appendix).[54] Consensus clustering and GSEA were implemented through GenePattern version 1.3.1 (Cambridge, MA), whereas hazard ratios were calculated using the statistical package SPSS version 11.0.1 (SPSS Inc, Chicago, IL). [55] All other analyses and graphs were performed using the R statistical programming language version 1.9.0 (Vienna, Austria) and Bioconductor version 1.4 (Seattle, WA).[56]

RESULTS

Demographic and Sample Characteristics

After exclusion of ineligible patients and array-based quality filtering, 31 Stanford, 72 Michigan, and 128 Dana-Farber adenocarcinomas were available for analysis. Examination of the available clinical covariates demonstrated the cohorts to be of a similar composition overall, although missing data precluded a thorough evaluation of the Stanford samples (Table 1). The distribution of age, smoking, sex, and BAC was remarkably similar for the Dana-Farber and Michigan cohorts. There was a trend toward a difference in stage distribution, with 79% of Michigan versus 69% of Dana-Farber samples with stage I or II disease ($P = .12$). Similarly, *K-ras* mutants were more common in the Michigan group (46% *v* 34%; $P = .10$). The most striking difference between the Michigan and Dana-Farber samples was the percentage of well-differentiated tumors (28% *v* 14%; $P = .02$). Also differing by cohort was the strategy by which adenocarcinoma samples were assigned to a subtype in the initial reports of the data. For example, in the Michigan scheme, every patient was slotted to one of three subtypes, whereas the Dana-Farber and Stanford groups left many samples unassigned. Similarly investigators differed in criteria for tumor inclusion in their respective studies. For example, to enrich for tumor-specific RNA, Michigan samples were selected to contain more than 70% tumor nuclei and exclude extensive fibrosis and inflammation. In contrast, the Dana-Farber set included samples with a minimum of 30% tumor nuclei, with estimated percentage tumor recorded in most cases. The difference in inclusion criteria introduces the possibility that clinically and biologically meaningful differences in the cohorts may have been introduced because approximately half of Dana-Farber tumors were composed of samples with less than 70% tumor nuclei. Selection of samples by percent tumor nuclei appears likely to account

Table 1. Patient Demographics and Tumor Characteristics by Data Source

Characteristic	Stanford University		University of Michigan		Dana-Farber Cancer Institute	
	No.	%	No.	%	No.	%
No. of patients	31		72		128	
Sex*						
Male	NA		30	42	47	41
Female	NA		42	58	67	59
Median age, years	NA		63.3		64.1	
Bronchioalveolar histology	NA			24		22
Nonsmoker	NA			10		10
Smoking < 10 years	NA			16		16
Stage*						
Ia	4		36		35	
Ib	4 (2)†		21		39	
IIa	1 (1)†		0		3	
IIb	1		0		19	
IIIa	6		12		7	
IIIb	0		3		3	
IV	9 (2)†		0		6	
Differentiation*						
Well	1	4	20	28	15	14
Moderate	14	54	34	47	66	62
Poor	11	42	18	25	26	24
EGFR mutation*	NA		NA		14 of 114	11
K-ras mutation*	NA		33 of 71	46	29 of 86	34
Published No. of adenocarcinoma clusters‡	3		3		4	
Clusters names and No. assigned to each as presented in original published reports§	A1 = 15 A2 = 6 A3 = 5 Unnamed cluster associated with large-cell adenocarcinoma = 5		1 = 17 2 = 35 3 = 20 All samples assigned to a cluster		C1 = 10 C2 = 12 C3 = 15 C4 = 15 Unnamed = 76	
Inclusion/exclusion criteria						
Histology	No restriction, any available lung tumor		Only adenocarcinoma, no adenosquamous, squamous, or other histology		No restriction, any available lung tumor	
Tumor % criteria	NA		70% minimum		30%, tumor minimum, verified by 2 pathologists	
Necrosis criteria	NA		NA		< 40% necrosis	
Fibrosis and inflammation	NA		"Extensive" fibrosis and inflammation excluded		Fibrosis and inflammation allowed	
Tissue source	Tumor bank		Single institution		Multiple tumor banks	
Treatment	NA		Stage I patients, resection and intrathoracic nodal sampling and no other treatments; stage III patients received surgical resection plus chemotherapy and radiotherapy		NA	

Abbreviation: NA, not available.
*Numbers do not sum to total because of missing data.
†Value in parentheses indicates number of samples missing survival data.
‡Number of subtypes reported in the original published reports.
§There is no implied association by row order. For example, A1, 1, and C1 are not assumed to represent the same cluster.

for differences in tumor grade seen between the cohorts (see Supplementary Data).

Gene Selection

The majority of excluded genes were ineligible because of absence on one or more of the three array platforms. An additional 40% of genes were discarded after being flagged as poorly measured by the SMD server. Of the remaining 2,848 genes, 90% (2,553) were reliable by the IC method and were used for additional analysis. A flow chart is provided in the Supplementary Data to document reasons for gene inclusion/exclusion in the current study.

Consensus Clustering: Identification of Bronchioid, Squamoid, and Magnoid Adenocarcinoma Subtypes

The identification of adenocarcinoma subtypes by hierarchical consensus clustering is shown in Figure 1, with three tumor subtypes suggested as optimum in each of the three cohorts. The choice of three clusters was confirmed using nonnegative matrix factorization–based consensus clustering (see Supplementary Data). Accordingly, within each cohort every sample was assigned exclusively to one of three subtypes defined by the consensus clusters. The nine centroids generated in this manner (one for each subtype in each dataset), as well as 10

reference centroids (three NL, two SQ, two SCLC, one LCLC, and two BAC), were evaluated for their pair-wise correlations across the 2,553 reliable genes using hierarchical agglomerative clustering (Fig 2). All centroids of similar histology, including NL, SQ, SCLC, and BAC, each derived from a different data source and array platform, demonstrate high correlation in the branched dendrogram. Similarly, adenocarcinoma subtype centroids demonstrate a strong cross-platform pattern of correlation in the following manner. Each dataset contributed one centroid to a dendrogram branch associated with the BAC centroids, thereby suggesting the cluster name bronchioid. Similarly, each dataset contributed a squamoid adenocarcinoma centroid to a dendrogram branch highly correlated with a SQ centroid. The remaining three adenocarcinoma centroids correlated best with the LCLC, offering the remaining centroid name of magnoid (from the Latin *magnus*, meaning "large"), although we note that the Michigan-derived centroid had an overall lower correlation. The results of tumor subtyping by consensus clustering were compared with results proposed in the original reports of these data in the Supplementary Data. The mapping of consensus clusters to those originally reported documents clear concordance in every case; it also highlights complex idiosyncrasies that impede a direct comparison of nonstandardized clustering.

Clinical and Biologic Correlates of Adenocarcinoma Subtypes

The adenocarcinoma subtypes were characterized by the available clinical and phenotypic data (Table 2). The subtype prevalence was similar across cohorts, with bronchioid and squamoid tumors comprising each around 33% to 52% of samples; the magnoid type comprised a minority at 10% to 26%. The percent tumor nuclei by subtype was highest in the bronchioid group and lowest in the squamoid group. In all three datasets, the squamoid subtype contained a higher percentage of poorly differentiated tumors than the bronchioid adenocarcinomas. Figure 2 suggests by the branch lengths of the dendrogram that the squamoid and magnoid clusters are more closely related to each other than either is to the bronchioid cluster. It is likely that this relationship is at least in part related to the properties they share of overall higher tumor grade and lower percentage tumor nuclei.

Adenocarcinoma subtype did not correlate clearly with stage in any of the datasets. All but one tumor with mucin was found in the bronchioid cluster. Clear cell histology was noted in four samples, none of which were of the bronchioid subtype. Interestingly, there was an over-representation of females, nonsmokers, and BAC histology in the bronchioid relative to the squamoid adenocarcinomas. Of all samples with any BAC histologic features reported in the pathologist's description, 75% fell within the bronchioid cluster. Accordingly, the highest percentage of epidermal growth factor receptor (*EGFR*) mutations was found in the bronchioid subtype (15%), with only one of 33 magnoid samples having an *EGFR* mutation. The single mutation found in the magnoid subgroup occurred in an extracellular domain of the gene not associated with responsiveness to EGFR inhibitors (unpublished data). Although the χ^2 *P* value failed to meet statistical testing for a difference in proportion of *EGFR* mutation by tumor subtype ($P = .21$), a trend was noted for the comparison of bronchioid versus magnoid ($P = .08$). Moreover, although not statistically significant, increased frequency of mutated K-ras was noted in the squamoid subtype relative to the bronchioid (30% *v* 37%; $P = .3$).

Kaplan-Meier curves were generated to assess differences in survival by adenocarcinoma subtype (Fig 3). Only the Dana-Farber group had sufficient follow-up and numbers of events to calculate curves for stage I and II patients. In these patients, the squamoid and magnoid subtypes demonstrated significantly shorter survival

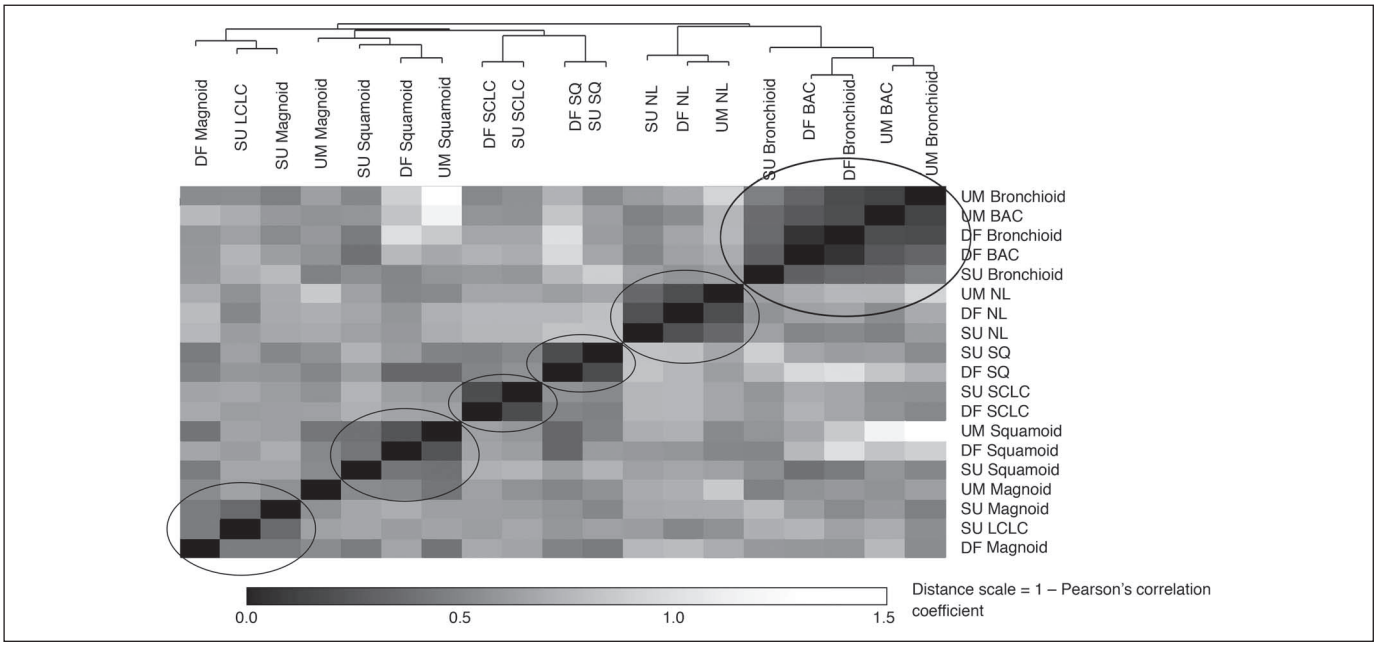

Fig 2. Hierarchical clustering of centroids derived from three independent gene expression datasets. At the intersection of each column and row in the figure is a pixel, the intensity of which is a measure of the distance (defined as 1 − Pearson's correlation coefficient) between the centroids named by the intersecting column and row (see text). DF, Dana-Farber Cancer Institute; SU, Stanford University; LCLC, large-cell lung cancer; UM, University of Michigan; SCLC, small-cell lung cancer; SQ, squamous cell lung carcinoma; NL, normal lung; BAC, bronchioalveolar carcinoma.

Table 2. Patient Demographics and Tumor Characteristics by Data Source and Cluster Assignment

Characteristic	Stanford University			University of Michigan			Dana-Farber Cancer Institute		
	B	S	M	B	S	M	B	S	M
No.	16	11	4	32	33	7	53	42	33
% of total by data source	52	35	13	44	46	10	41	33	26
Mean % tumor	NA	NA	NA	NA	NA	NA	72	57	68
Sex*									
Male	NA	NA	NA	12	15	4	18	17	12
Female	NA	NA	NA	20	18	3	33	13	21
Median age, years	NA	NA	NA	65.2	63.2	65.6	64	65.5	65
Bronchioalveolar histology, %	NA	NA	NA	34	12	28	39	12	3
Nonsmoker, %	NA	NA	NA	9	6	29	11	5	9
Stage*									
Ia	3	1	0	15	17	4	17	6	12
Ib	3	1	0	10	8	3	20	12	7
IIa	1	1	0	0	0	0	1	1	1
IIb	0	0	0	0	0	0	8	5	6
IIIa	2	4	0	5	7	0	2	0	5
IIIb	0	0	0	2	1	0	1	1	1
IV	5	3	1	0	0	0	1	4	1
Differentiation*									
Well	1	0	3	12	5	3	11	4	0
Moderate	10	4	0	14	19	1	33	17	16
Poor	3	7	0	6	9	3	3	7	16
Histologic features									
Clear cell	NA	NA	NA	0	3	0	0	0	1
Papillary	NA	NA	NA	4	3	0	4	0	2
Mucin	NA	NA	NA	6	1	0	2	0	0
% EGFR mutation*	NA	NA	NA	NA	NA	NA	15	12	3
% K-ras Mutation*	NA	NA	NA	42	52	43	23	26	18

Abbreviations: B, bronchioid adenocarcinoma subtype; S, squamoid adenocarcinoma subtype; M, magnoid adenocarcinoma subtype; NA, not available.
*Numbers do not sum to total due to missing data.

compared with the bronchioid tumors, with hazard ratios (HRs) of 3.6 ($P = .01$) and HR 3.0 ($P = .04$), respectively. After stratifying by stage, we evaluated all clinical covariates available in these data by multivariate Cox proportional hazards modeling for association with survival, including age, differentiation, sex, smoking status, BAC histology, K-ras mutations status, and *EGFR* mutation status. In both the univariate and multivariate analysis, only tumor subtype, age, and differentiation were significantly associated with survival. Strikingly, in advanced and nonsurgical disease (stages III and IV), the survival advantage is reversed with a trend toward improved survival in the squamoid subtype relative to the magnoid (HR, 0.32; $P = .03$) and bronchioid subtypes (HR, 0.58; $P = .2$). There were too few samples to perform meaningful multivariate adjustment in advanced-stage patients. Evaluating all eligible patients without regard to stage, survival was worse in magnoid tumors compared with both bronchioid (HR, 1.7; $P = .04$.) and squamoid (HR, = 1.6; $P = .10$) tumors. Absent consideration of stage, survival in the squamoid versus bronchioid tumors essentially is identical (HR, 1.1; $P = .70$).

Differential survival by tumor subtype was confirmed in two independent cohorts of early-stage lung adenocarcinoma patients treated by surgery alone using identical methods to those described above (see Fig 4 and Supplementary Data). The first was a group of 41 patients treated at Duke University whose tumors were assayed using the Affymetrix u1332plus GeneChip with approximately 47,000 transcripts.[57] As in the Dana-Farber example, the squamoid and magnoid subtypes demonstrated inferior survival compared with the bronchioid (HR, 8.1; $P < .001$ and HR, 9.7; $P < .001$, respectively). The second cohort of 86 patients was constructed at the University of British Columbia for the purpose of evaluating 18 immunohistochemical markers in non–small-cell lung cancer patients using a tissue microarray format and has been described in detail elsewhere.[58] These data were particularly interesting because the markers measured proteins as opposed to gene expression. Yet again, the results were consistent, with the squamoid and magnoid subtypes demonstrating clear trends toward inferior survival compared with the bronchioid samples (HR, 2.7; $P = .06$ and HR, 2.2; $P = .16$, respectively).

Incidence and site of distant recurrence were available for early-stage tumors from the Dana-Farber cohort. Of 74 patients with stage I disease, 28 patients (38%) had a recurrence reported in the study period. Both the pattern and rate of recurrence varied by tumor subtype, however, with 27% of patients with bronchioid, 61% of patients with squamoid, and 37% of patients with magnoid subtypes reporting a recurrence ($P = .04$). Interestingly, five of six patients with bronchioid tumors and distant metastases reported bone involvement, representing 63% of all bone recurrences in these data. Finally, five of nine patients with squamoid tumors and distant metastases reported brain involvement, representing 71% of all brain recurrences.

Genes Associated With Subtypes

For each dataset, all possible one group versus all groups, and all pair-wise comparisons were evaluated by the SAM methodology, generating the 36 gene lists described in Table 3. The genes corresponding

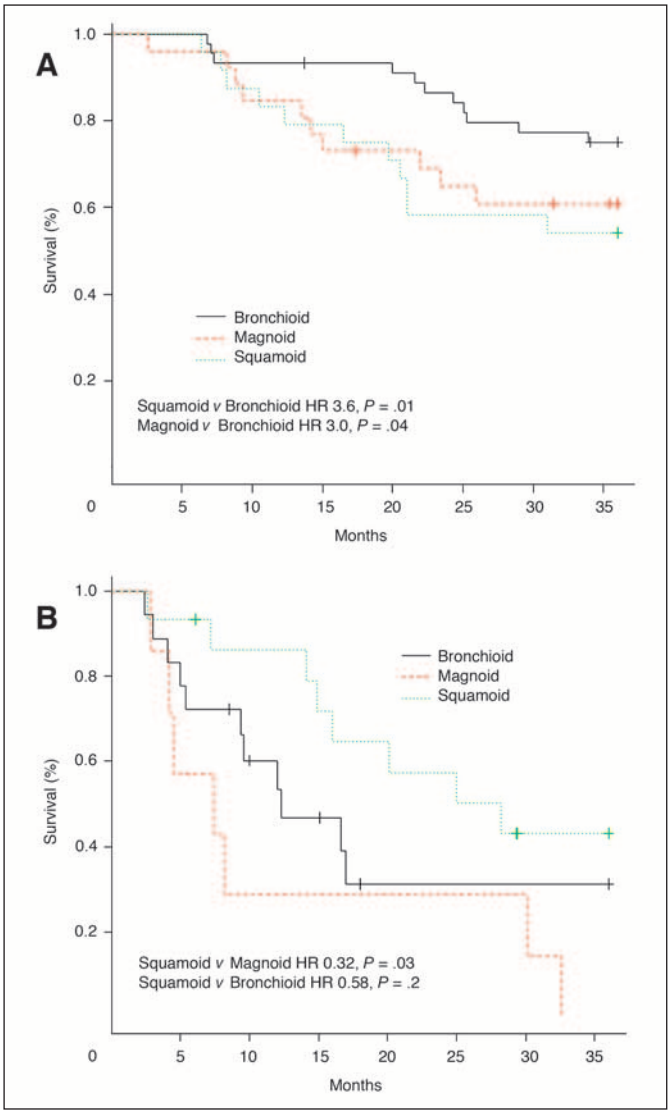

Fig 3. Survival by stage and adenocarcinoma subtype. Survival in patients with (A) stages I and II and (B) stages III and IV adenocarcinoma of the lung as a function of adenocarcinoma subtype derived from gene expression arrays. HR, hazard ratio.

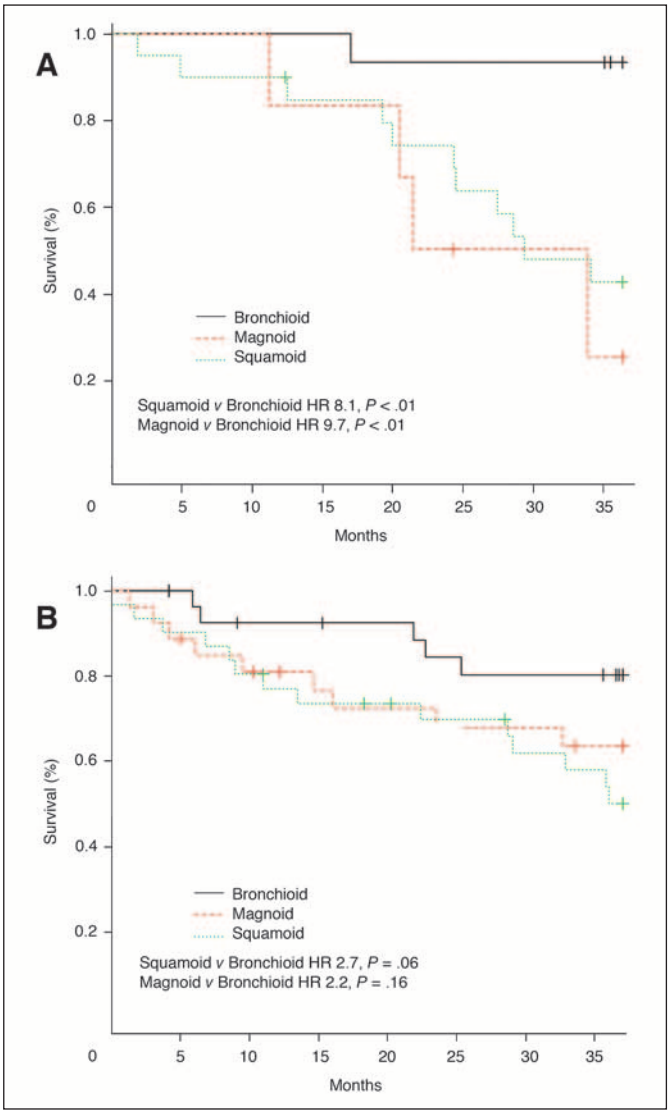

Fig 4. Survival by adenocarcinoma subtype in two independent validation cohorts. (A) Survival in 41 early-stage lung adenocarcinoma patients as a function of expression microarray–based tumor subtype. (B) Survival in 85 surgically resected lung adenocarcinoma patients as a function of adenocarcinoma subtype. Subtypes were derived based on 18 immunohistochemical markers performed on paraffin-embedded tissue using a tissue microarray system. HR, hazard ratio.

to each cell of the table are available in the Supplementary Data. Of the 2,553 reliable genes, 1,066 (42%) were selected by SAM at least once. As expected, the number of differentially expressed genes correlated with the numbers of patients in the cohort. Of interest, considerably fewer genes per hypothesis tested were identified in the Stanford group even after accounting for cohort size; this result probably reflects technical features of gene expression measurement in the Stanford microarray platform. Gene lists derived from the Michigan data were comparable in length to those from the Dana-Farber group, with the exception of those for the magnoid subtype, which were shortened and had higher FDRs.

The SAM-generated lists were examined for concordance (Table 4). In 28 of 36 comparisons, the concordance across gene lists was greater than expected by chance. Of those for which concordance was not greater than expected by chance, five of eight involved the Michigan magnoid subtype. GSEA was performed to test the statistical significance of SAM-generated gene lists as independent predictors of tumor subtypes across studies (Table 5). Of the 72 gene lists validated, there was evidence supporting cross-platform validation in 59. Of the 13 lists that failed to validate by these criteria, nine involved the Michigan magnoid subtype, demonstrating its weak signature in these data.

Biologic Pathways of Tumor Subtypes

Although an explicit evaluation of the tumor subtype biology is outside the scope of this article, a brief consideration clearly is warranted (Table 6). Bronchioid tumors were dominated by a program of growth, development, differentiation, and survival genes. Defining genes of the squamoid tumors stem from a dramatically different set of tumor processes, including angiogenesis such as hypoxia-inducible factor-1-alpha, transforming growth factor beta pathway genes, and

Table 3. Number of Genes Discriminating Adenocarcinoma Subtypes by Sample Source

Subtype	Dana-Farber Cancer Institute		University of Michigan		Stanford University	
	No. of Genes	False Discovery Rate*	No. of Genes	False Discovery Rate	No. of Genes	False Discovery Rate
Bronchioid *v* all†	323	0.003	265	0.001	14	0.014
All *v* bronchioid‡	734	0.003	806	0.001	46	0.001
Bronchioid *v* squamoid	280	0.001	277	0.001	14	0.030
Bronchioid *v* magnoid	156	0.002	35	0.045	63	0.001
Squamoid *v* all	461	0.001	718	0.001	42	0.001
All *v* squamoid	189	0.001	228	0.001	19	0.297
Squamoid *v* bronchioid	634	0.001	797	0.001	55	0.001
Squamoid *v* magnoid	222	0.001	43	0.001	45	0.001
Magnoid *v* all	281	0.002	15	0.060	18	0.005
All *v* magnoid	120	0.001	123	0.639	74	0.001
Magnoid *v* bronchioid	448	0.002	91	0.001	16	0.009
Magnoid *v* squamoid	181	0.001	59	0.053	11	0.012

*See Methods.
†Can be interpreted as number of genes with increased expression in bronchioid relative to all other samples.
‡Can be interpreted as number of genes with decreased expression in bronchioid relative to all other samples.

the WNT signaling cascade. Magnoid tumors demonstrate a pattern of gene expression associated with a distinct set of pathways, primarily inflammation, cytoskeleton, metabolism, and proliferation. In addition, of clinical interest we note that the three subtypes differ with respect to a number of putative markers of cancer chemotherapy and radiation treatment. Of particular note, the bronchioid subtype is associated with the majority of genes associated with cisplatin resistance. A more in-depth review of these genes is provided in the Supplementary Data.

DISCUSSION

The hypothesis tested for the first time in the current work is that lung adenocarcinoma subtypes defined by gene array analysis are reproducible and clinically relevant. The adenocarcinoma subtypes we report were identified in an unbiased, independent, and objective manner, and are distinct in cross-platform validation by correlation with expression patterns from recognized lung tumor histologic subtypes (BAC, LCLC, SQ, and SCC). Furthermore, reproducible subtypes can be identified through the use of centroids even in the absence of a gold standard reference such as a molecular marker or an a priori predictive gene list. Tumor subtypes were named according to overall similarity of gene expression patterns across hundreds or thousands of genes to easily recognizable morphologic lung cancer variants. This naming choice emphasizes the view that the tumor subtypes are not dependent on identification of a fixed set of genes, specific analytic method, or microarray platform, and allows future investigators to establish a common reference point lacking in this heterogeneous disease.

Most notably, the three independent datasets each produced clear pictures of the bronchioid and squamoid adenocarcinoma subtypes. With regard to their clinical features, bronchioid tumors were more likely to be from nonsmoking females with BAC histology and

Table 4. Gene List Overlap by Sample Source and Cluster

Subtype	Dana-Farber Cancer Institute/University of Michigan		Stanford University/Dana-Farber Cancer Institute		University of Michigan/Stanford University	
	Observed Count	Expected Count Due to Chance Alone	Observed Count	Expected Count Due to Chance Alone	Observed Count	Expected Count Due to Chance Alone
Bronchioid *v* all	111	33	3	1	4	1
All *v* bronchioid	457	231	13	13	11	14
Bronchioid *v* squamoid	101	30	2	1	2	1
Bronchioid *v* magnoid	8	2	6	3	1	0
Squamoid *v* all	272	129	8	7	16	11
All *v* squamoid	58	16	3	1	0	1
Squamoid *v* bronchioid	413	197	7	13	18	17
Squamoid *v* magnoid	5	3	14	3	1	0
Magnoid *v* all	0	1	3	2	0	0
All *v* magnoid	5	5	11	3	5	3
Magnoid *v* bronchioid	28	16	3	2	0	0
Magnoid *v* squamoid	4	4	0	0	1	0

Table 5. Gene Set Enrichment Analysis

Subtype	Dana-Farber Cancer Institute/ University of Michigan		Dana-Farber Cancer Institute/ Stanford University		Stanford University/ University of Michigan		Stanford University/ Dana-Farber Cancer Institute		University of Michigan/Dana-Farber Cancer Institute		University of Michigan/Stanford University	
	ES	*P*	ES	*P*	ES	*P*	ES	*P*	ES	*P*	ES	*P*
Bronchioid *v* all	0.4*	0	0.1†	.15	0.3‡	.5	0.5*	.05	0.5*	0	0.1†	.6
All *v* bronchioid	0.3*	0	0.1†	.2	0.1‡	.95	0.2‡	.7	0.3*	0	0.1†	.55
Bronchioid *v* squamoid	0.5*	0	0.1†	.15	0.2‡	.65	0.2‡	.8	0.4*	0	0.1†	.1
Bronchioid *v* magnoid	0.3†	.2	0.2†	.2	−0.1§	.7	0.3*	.05	0.4*	0	−0.2§	.1
Squamoid *v* all	0.4*	0	0.1*	0	0.2‡	.5	−0.1§	.99	0.3*	0	0.1†	.25
All *v* squamoid	0.4*	0	0.2†	.2	0.2‡	.6	0.2‡	.4	0.4*	0	0.1†	0.1
Squamoid *v* bronchioid	0.4*	0	0.1†	.25	0.2‡	.5	−0.1§	.99	0.3*	0	0.1‡	0.4
Squamoid *v* magnoid	0.1†	.15	0.2*	0	0.2‡	.65	0.3†	.15	0.2†	.1	0.1‡	.85
Magnoid *v* all	0.1‡	.8	0.3*	0	0.1‡	.9	0.3†	.25	−0.3‖	.05	−0.1§	.99
All *v* magnoid	−0.1§	.9	0.2†	.1	0.2‡	.7	0.3*	0	−0.1§	.5	−0.1§	.2
Magnoid *v* bronchioid	0.2†	.1	0.1†	.25	−0.2§	.35	0.4†	.25	0.2†	.1	−0.1§	.9
Magnoid *v* squamoid	0.2†	.1	0.2*	.05	0.1‡	.99	−0.2§	.99	0.1‡	.75	−0.2§	.2

NOTE. The column heading names the dataset that generated the gene list, followed by the cohort in which the list was validated. The ES sign (+ or −) denotes the direction of correlation of the gene list with the tumor subtype distinction named in the row (see Methods and Appendix).
Abbreviation: ES, enrichment score.
*Evidence for validation, ES score positive, permutation *P* significant.
†Evidence for validation, ES score positive, permutation *P* trending toward significance (.1 to .25).
‡Evidence for validation, ES score positive, permutation $P > .25$.
§Evidence against validation, ES score negative, permutation *P* not significant.
‖Evidence against validation, ES score negative, permutation *P* significant.

contain mutations of the *EGFR* gene. Patients with these tumors demonstrated significantly improved survival compared with other tumor subtypes in early-stage disease, but poorer survival in late-stage disease. Improved survival for early-stage bronchioid patients may be due to their lower rate of distant metastases compared with the other tumor subtypes. When bronchioid tumors did metastasize, the recurrence tended to occur in bone. Why bronchioid tumors might fare worse in advanced disease is unclear from these data, although genes

Table 6. Genes by Tumor Subtype and Biologic Category

Bronchioid
Growth, development, differentiation, and survival/antiapoptosis
Differentiation: retinoid X receptors (alpha, beta, and gamma), *RARG*, *ABCA4*, *THRA*, *TRIP3*
Growth and development: *DLX4*, *IRX5*, *LHX2*, *CPDP1*, *ARVCF* (velocardiofacial syndrome), *PAX3* (Waardenburg's syndrome 1), *MSX2* (craniosynostosis), faciogenital dysplasia (Aarskog-Scott syndrome), *RUNX2* (cleidocranial dysplasia), *UBE3A* (Angelman syndrome)
ETS genes: *CDK10*, *ETV3*, *ETV4*, *ELK1*, FOS-like antigen 2
JAK/STAT and antiapoptosis genes: *PIK3R2*, *PIK3CD*, *STAT5B*, *IL6R*, *CCND2* (decreased), *p21* (decreased)
Extracellular matrix and matrix metalloproteinases: *ST3GAL2*, *ST3GAL4*, *ALG3*, *CSPG4*, *MGAT3*, *SDC2*, *MMP15*, *MMP17*, *ADAMS11*
Type II pneumocyte: *MUC1*, *ABCA3*
Cisplatin resistance, radiation, and DNA repair: *ERCC2*, *XRCC1*, *XRCC5*, *WWP2*, *LIG3*
Squamoid
WNT-HDAC2,APC (decreased), *MLLT3*, *WNT5A*, *CCND2*, *ADAM9* and *ADAM10*, *TFRC*, *BLMH*
Angiogenesis: *TCEB1*, *VHL* (decreased), *HIF1A*, *ATR*, *RPS6KA1*, *CREBBP*
Squamous cell markers/differentiation: *SART3*, *CKS1B*, *ERK3*, *ADAM9*, *CD24*, *CXCR4*, *PML* (decreased), *XBP1*, *SMAD1*, *SMAD2*, *SMAD4*, *BMPR*, *BMP6*, *ID1*, *ID2*, *ID3*
Compliment: *CD55*, *CD46*, and *CD59*
Zellweger's syndrome: *SCP2*, *PXMP3*
Translation
RNA helicases: *DEAD* Box polypeptides 1, 5, 18, 21, and 48
RNA Polymerase II: *TAF7*, *TAF9*, *TAF11*, *SKP1A*, *GTF2F1*
Chemotherapy targets: *MTHFD2*, *MTHFD1*, *DTYMK*, *DCK*, *FOLR1*, *CYR61*, *BLMH*, *CLU* (decreased), *EHHX1*, *EPHX2*
Magnoid
Inflammatory genes: *ILF3*, *TNFAIP2*, *PLAUR*, *IRAK1*, *IL15RA*, *FCGR2B*, *FCGR3A*, *MCM3*, *ANXA1*, *IFI35*, *IFRD1*
Cytoskeleton: *TUBB5*, *PIP5K2A*, *LIMS1*, *ADD2*, *TROAP*, *TGFBI*, *CDKN2A*, *FLNA*, *TUBG1*, *TPM2*, *MAP4*, *SNTA1*, *EXOSC10*, *RSN*, *PDLIM4*, *ARPC1B*, *ARPC2*, *VIM*, *CKAP4*, *PLOD1*, *PLOD2*, *DAG1*, *ICAM1*
Hematopoietic markers: *MMD*, *HEM1*
Lung/epithelial markers: *DNAJA1*, *EMP3*, *MMP10*, *ATF4*, *FUS*, *EZH2*, *NME1*, *ST3GAL3*, *PRKCSH*, *ERCC3*
Proliferation: *MKI67* (Ki-67), *PCNA*, *CBX3*, *EIF2S1*, *EIF5*, *EIF3S2*
Genes associated with chemotherapy targets: *FNTA*, *FDPS*, *TAP1*(*MDR*/*TAP*), *TYMS*, *TK1*, *TOP2A*, *TOPBP1*
Neuroendocrine: ADM

defining the bronchioid subtype were more likely to be those correlated with chemotherapy and radiation resistance. Bronchioid tumors were of overall lower tumor grade, tended to demonstrate markers of type II pneumocyte differentiation, and stain positively for mucin production. In contrast, squamoid tumors were more likely from male smokers in tumors with *K-ras* gene mutation. Patients with squamoid tumors fared significantly worse that those with the bronchioid subtype in early-stage disease, but better in advanced disease. The poor prognosis in earlier stages is likely due to a tendency to metastasize early, including a higher likelihood of brain involvement. Squamoid adenocarcinomas were more likely to be moderately or poorly differentiated and to be associated with genes most commonly associated with squamous cell carcinoma. Gene list predictors of squamoid and bronchioid subtypes were generated independently for each subtype in each cohort, and in every case these were validated by the GSEA and centroid methodologies.

A third group, the magnoid subtype, was also selected in each of the three independent cohorts by the objective method we describe. Magnoid tumors were the most infrequent, ranging from 10% in the Michigan cohort to 26% in the Dana-Farber samples. One of the most pronounced characteristics of the magnoid subtype was the strong inflammatory signature. Presumably, because the exclusion criteria for the Michigan cohort included significant numbers of inflammatory cells, this would reduce the percentage of magnoid tumors in this cohort. As a result, although all three cohorts detect the magnoid cluster, both the gene expression and clinical profile are less distinct than for the bronchioid and squamoid tumors, although the overall poor prognosis of the group was statistically significant in advanced disease.

The expression patterns defining the tumor subtypes presented are not subtle statistical phenomena dependent on a handful of predictive genes. Forty percent of all reliable genes in the dataset were predictive of at least one adenocarcinoma subtype using the criteria we established in the Methods section. Amazingly, the expression signature extends beyond even these 1,066 genes selected by SAM. When all genes selected as predictors of the tumor subtypes are excluded and the current analysis repeated, we obtain essentially identical results (data not shown). In other words, many genes failing to meet significance testing will contribute signal to the tumor subtype identity by the centroid method analysis.

By focusing on standardized and unbiased methods, we essentially have excluded the possibility that adenocarcinoma subtypes are the result of chance, noise, artifact, or analytic method. None of the analytic parameters, including sample selection, gene selection, optimal cluster number, and cluster assignment, were optimized with regard to the study outcome. In each case, analytic methods were based on a priori biologic and statistical considerations. Although unrecognized technical artifacts can drive clustering patterns in a single dataset, it is unlikely that similar effects would be present in multiple cohorts using different assay platforms, as was the case here. It is even less likely that spurious clusters would correlate with the constellations of clinical features across three datasets in the manner described in this study. In addition, our standardized analysis agreed well with the previously published results, but also clarified the findings for meaningful comparison that would not otherwise be possible. Finally, we demonstrate the ability to evaluate tumor subtypes with confidence and ease in a platform-independent manner, as we did with the expression arrays from Duke and the tissue microarrays from the University of British Columbia.

Although a specific discussion of genes and biologic pathways is beyond the scope of this work, all of the data, including the lists of genes associated with each of the subtypes, are available in the Supplementary Data. Regarding the cancer pathways associated with the tumor subtypes, our results mirror those of the previously published reports.[20-22] The importance of tumor subtyping is clear even in the absence of a complete biologic understanding. Tumor subtype is suggested as a proxy for at least one important mutation (*EGFR*), with none of the magnoids demonstrating the clinically meaningful finding. It is likely that other important genomic events are conferred by tumor subtype membership, including specific chromosomal abnormalities (results not shown).

The main focus of this analysis was the validation of adenocarcinoma subtypes derived from clustering of expression profiles. We chose to validate three subtypes, given that this was the number suggested by consensus clustering. It is striking that, in the face of what is considered a heterogeneous tumor, three clusters emerged consistently, suggesting that a molecular taxonomy could be proposed that is simple, reproducible, and complementary to light microscopy alone. We do not exclude the possibility that additional tumor subtypes might be described if the sample set were larger or of a different composition. A group of investigators funded under the National Institutes of Health's Directors Challenge Program recently has completed processing of several hundred new lung cancer expression arrays. It is hoped that the validation of lung adenocarcinoma subtypes in the current report in conjunction with these new data will expedite more reliable classification of the heterogeneous group of tumors currently known most frequently as NSCLC.

REFERENCES

1. Parkin DM: Global cancer statistics in the year 2000. Lancet Oncol 2:533-543, 2001

2. Travis WD, Sobin LH: Histological Typing of Lung and Pleural Tumours (ed 3). New York, NY, Springer-Verlag, 1999

3. Petrovich Z, Mietlowski W, Ohanian M, et al: Clinical report of the treatment of locally advanced lung cancer. Cancer 40:72-77, 1977

4. Morstyn G, Ihde DC, Lichter AS, et al: Small cell lung cancer 1973-1983: Early progress and recent obstacles. Int J Radiat Oncol Biol Phys 10:515-539, 1984

5. Lynch TJ, Bell DW, Sordella R, et al: Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N Engl J Med 350:2129-2139, 2004

6. Paez JG, Janne PA, Lee JC, et al: EGFR mutations in lung cancer: Correlation with clinical response to gefitinib therapy. Science 304:1497-1500, 2004

7. Chemotherapy in non-small cell lung cancer: A meta-analysis using updated data on individual patients from 52 randomised clinical trials: Non-Small Cell Lung Cancer Collaborative Group. BMJ 311:899-909, 1995

8. Yamamoto S, Sobue T, Yamaguchi N, et al: Reproducibility of diagnosis and its influence on the distribution of lung cancer by histologic type in Osaka, Japan. Jpn J Cancer Res 91:1-8, 2000

9. Aida S, Shimazaki H, Sato K, et al: Prognostic analysis of pulmonary adenocarcinoma subclassification with special consideration of papillary and bronchioloalveolar types. Histopathology 45:468-476, 2004

10. Sorensen JB, Hirsch FR, Gazdar A, et al: Interobserver variability in histopathologic subtyping and grading of pulmonary adenocarcinoma. Cancer 71:2971-2976, 1993

11. Ghandur-Mnaymneh L, Raub WA Jr, Sridhar KS, et al: The accuracy of the histological classification of lung carcinoma and its reproducibility: A study of 75 archival cases of adenosquamous carcinoma. Cancer Invest 11:641-651, 1993

12. Meyerson M, Hayes DN: Microarray approaches to gene expression analysis, in Tsongalis

GJ, Coleman WB (eds): Molecular Diagnostics: For the Clinical Laboratorian (ed 2). Totowa, NJ, Humana Press, 2005, pp 121-148

13. Alizadeh AA, Ross DT, Perou CM, et al: Towards a novel classification of human malignancies based on gene expression patterns. J Pathol 195: 41-52, 2001

14. 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 100:8418-8423, 2003

15. Cleator S, Ashworth A: Molecular profiling of breast cancer: Clinical implications. Br J Cancer 90:1120-1124, 2004

16. Takeuchi T, Tomida S, Yatabe Y, et al: Expression profile-defined classification of lung adenocarcinoma shows close relationship with underlying major genetic changes and clinicopathologic behaviors. J Clin Oncol 24:1679-1688, 2006

17. Jiang H, Deng Y, Chen HS, et al: Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. BMC Bioinformatics 5:81, 2004

18. Parmigiani G, Garrett-Mayer ES, Anbazhagan R, et al: A cross-study comparison of gene expression studies for the molecular classification of lung cancer. Clin Cancer Res 10:2922-2927, 2004

19. Yamagata N, Shyr Y, Yanagisawa K, et al: A training-testing approach to the molecular classification of resected non-small cell lung cancer. Clin Cancer Res 9:4695-4704, 2003

20. Bhattacharjee A, Richards WG, Staunton J, et al: Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A 98:13790-13795, 2001

21. Garber ME, Troyanskaya OG, Schluens K, et al: Diversity of gene expression in adenocarcinoma of the lung. Proc Natl Acad Sci U S A 98:13784-13789, 2001

22. Beer DG, Kardia SL, Huang CC, et al: Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med 8:816-824, 2002

23. Bonner AE, Lemon WJ, Devereux TR, et al: Molecular profiling of mouse lung tumors: Association with tumor progression, lung development, and human lung adenocarcinomas. Oncogene 23:1166-1176, 2004

24. Borczuk AC, Gorenstein L, Walter KL, et al: Non-small-cell lung cancer molecular signatures recapitulate lung developmental pathways. Am J Pathol 163:1949-1960, 2003

25. Borczuk AC, Shah L, Pearson GD, et al: Molecular signatures in biopsy specimens of lung cancer. Am J Respir Crit Care Med 170:167-174, 2004

26. Gordon GJ, Jensen RV, Hsiao LL, et al: Using gene expression ratios to predict outcome among patients with mesothelioma. J Natl Cancer Inst 95:598-605, 2003

27. Hoang CD, D'Cunha J, Tawfic SH, et al: Expression profiling of non-small cell lung carcinoma identifies metastatic genotypes based on lymph node tumor burden. J Thorac Cardiovasc Surg 127: 1332-1342, 2004

28. Jones MH, Virtanen C, Honjoh D, et al: Two prognostically significant subtypes of high-grade lung neuroendocrine tumours independent of small-cell and large-cell neuroendocrine carcinomas identified by gene expression profiles. Lancet 363: 775-781, 2004

29. Kikuchi T, Daigo Y, Katagiri T, et al: Expression profiles of non-small cell lung cancers on cDNA microarrays: Identification of genes for prediction of lymph-node metastasis and sensitivity to anti-cancer drugs. Oncogene 22:2192-2205, 2003

30. McDoniels-Silvers AL, Stoner GD, Lubet RA, et al: Differential expression of critical cellular genes in human lung adenocarcinomas and squamous cell carcinomas in comparison to normal lung tissues. Neoplasia 4:141-150, 2002

31. Miura K, Bowman ED, Simon R, et al: Laser capture microdissection and microarray expression analysis of lung adenocarcinoma reveals tobacco smoking- and prognosis-related molecular profiles. Cancer Res 62:3244-3250, 2002

32. Nakamura H, Saji H, Ogata A, et al: cDNA microarray analysis of gene expression in pathologic Stage IA nonsmall cell lung carcinomas. Cancer 97:2798-2805, 2003

33. Pass HI, Liu Z, Wali A, et al: Gene expression profiles predict survival and progression of pleural mesothelioma. Clin Cancer Res 10:849-859, 2004

34. Pedersen N, Mortensen S, Sorensen SB, et al: Transcriptional gene expression profiling of small cell lung cancer cells. Cancer Res 63:1943-1953, 2003

35. Singhal S, Amin KM, Kruklitis R, et al: Alterations in cell cycle genes in early stage lung adenocarcinoma identified by expression profiling. Cancer Biol Ther 2:291-298, 2003

36. Singhal S, Amin KM, Kruklitis R, et al: Differentially expressed apoptotic genes in early stage lung adenocarcinoma predicted by expression profiling. Cancer Biol Ther 2:566-571, 2003

37. Sugita M, Geraci M, Gao B, et al: Combined use of oligonucleotide and tissue microarrays identifies cancer/testis antigens as biomarkers in lung carcinoma. Cancer Res 62:3971-3979, 2002

38. Tomida S, Koshikawa K, Yatabe Y, et al: Gene expression-based, individualized outcome prediction for surgically treated lung cancer patients. Oncogene 23:5360-5370, 2004

39. Virtanen C, Ishikawa Y, Honjoh D, et al: Integrated classification of lung tumors and cell lines by expression profiling. Proc Natl Acad Sci U S A 99:12357-12362, 2002

40. Wigle DA, Jurisica I, Radulovich N, et al: Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. Cancer Res 62:3005-3008, 2002

41. Wikman H, Kettunen E, Seppanen JK, et al: Identification of differentially expressed genes in pulmonary adenocarcinoma by using cDNA array. Oncogene 21:5804-5813, 2002

42. Xi L, Lyons-Weiler J, Coello MC, et al: Prediction of lymph node metastasis by analysis of gene expression profiles in primary lung adenocarcinomas. Clin Cancer Res 11:4128-4135, 2005

43. Endoh H, Tomida S, Yatabe Y, et al: Prognostic model of pulmonary adenocarcinoma by expression profiling of eight genes as determined by quantitative real-time reverse transcriptase polymerase chain reaction. J Clin Oncol 22:811-819, 2004

44. Gordon GJ, Richards WG, Sugarbaker DJ, et al: A prognostic test for adenocarcinoma of the lung from gene expression profiling data. Cancer Epidemiol Biomarkers Prev 12:905-910, 2003

45. Hu Z, Fan C, Oh DS, et al: The molecular portraits of breast tumors are conserved across microarray platforms. BMC Genomics 7:96, 2006

46. Irizarry RA, Hobbs B, Collin F, et al: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4:249-264, 2003

47. Gollub J, Ball CA, Binkley G, et al: The Stanford Microarray Database: Data access and quality assessment tools. Nucleic Acids Res 31: 94-96, 2003

48. Bloom G, Yang IV, Boulware D, et al: Multi-platform, multi-site, microarray-based human tumor classification. Am J Pathol 164:9-16, 2004

49. Wheeler DL, Church DM, Federhen S, et al: Database resources of the National Center for Biotechnology. Nucleic Acids Res 31:28-33, 2003

50. Cope L, Zhong X, Garrett E, et al: MergeMaid: R tools for merging and cross-study validation of gene expression data. Stat Appl Genet Mol Biol 3, 2004 (Article 29; Epub: October 31, 2004)

51. Monti S, Tamayo P, Mesirov J, et al: Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning 52:91-118, 2003

52. Brunet JP, Tamayo P, Golub TR, et al: Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A 101:4164-4169, 2004

53. Tusher VG, Tibshirani R, Chu G: Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 98:5116-5121, 2001

54. 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 102:15545-15550, 2005

55. Reich M, Liefeld T, Gould J, et al: GenePattern 2.0. Nat Genet 38:500-501, 2006

56. Gentleman RC, Carey VJ, Bates DM, et al: Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol 5:R80, 2004

57. Bild AH, Yao G, Chang JT, et al: Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 439:353-357, 2006

58. Au NH, Cheang M, Huntsman DG, et al: Evaluation of immunohistochemical markers in non-small cell lung cancer by unsupervised hierarchical clustering analysis: A tissue microarray study of 284 cases and 18 markers. J Pathol 204:101-109, 2004

Acknowledgment

We thank David Beer, PhD, Cheng Li, PhD, Mitchell Garber, PhD, Anil Potti, MD, and Robert Gentleman, PhD, for their discussions and advice on this project.

Appendix

The Appendix is included in the full-text version of this article, available online at www.jco.org. It is not included in the PDF version (via Adobe® Reader®).

Authors' Disclosures of Potential Conflicts of Interest

Although all authors completed the disclosure declaration, the following author or immediate family members indicated a financial interest. No conflict exists for drugs or devices used in a study if they are not being evaluated as part of the investigation. For a detailed description of the disclosure categories, or for more information about ASCO's conflict of interest policy, please refer to the Author Disclosure Declaration and the Disclosures of Potential Conflicts of Interest section in Information for Contributors.

Authors	Employment	Leadership	Consultant	Stock	Honoraria	Research Funds	Testimony	Other
Matthew Meyerson			Novartis			Novartis		

Author Contributions

Conception and design: D. Neil Hayes, Katsuhiko Naoki, Arindam Bhattacharjee, Matthew Meyerson
Financial support: Matthew Meyerson
Administrative support: Matthew Meyerson
Collection and assembly of data: D. Neil Hayes, C. Blake Gilks, Matthew Meyerson
Data analysis and interpretation: D. Neil Hayes, Stefano Monti, Giovanni Parmigiani, C. Blake Gilks, Katsuhiko Naoki, Arindam Bhattacharjee, Mark A. Socinski, Charles Perou, Matthew Meyerson
Manuscript writing: D. Neil Hayes, Stefano Monti, Giovanni Parmigiani, C. Blake Gilks, Katsuhiko Naoki, Mark A. Socinski, Charles Perou, Matthew Meyerson
Final approval of manuscript: D. Neil Hayes, Stefano Monti, Giovanni Parmigiani, Katsuhiko Naoki, Arindam Bhattacharjee, Mark A. Socinski, Charles Perou, Matthew Meyerson