Transcriptomic Analysis of Human Lung Development

Alvin T. Kho^{1*}, Soumyaroop Bhattacharya^{2*}, Kelan G. Tantisira^{3*}, Vincent J. Carey³, Roger Gaedigk⁴, J. Steven Leeder⁴, Isaac S. Kohane¹, Scott T. Weiss³, and Thomas J. Mariani²

¹Children's Hospital Informatics Program, Harvard-MIT Division of Health Sciences and Technology, Boston, Massachusetts; ²Division of Neonatology and Center for Pediatric Biomedical Research, University of Rochester, Rochester, New York; ³Channing Laboratory, Brigham and Women's Hospital, Boston, Massachusetts; and ⁴Division of Clinical Pharmacology and Medical Toxicology, Department of Pediatrics, Children's Mercy Hospitals and Clinics, Kansas City, Missouri

Rationale: Current understanding of the molecular regulation of lung development is limited and derives mostly from animal studies. **Objectives:** To define global patterns of gene expression during human lung development.

Methods: Genome-wide expression profiling was used to measure the developing lung transcriptome in RNA samples derived from 38 normal human lung tissues at 53 to 154 days post conception. Principal component analysis was used to characterize global expression variation and to identify genes and bioontologic attributes contributing to these variations. Individual gene expression patterns were verified by quantitative reverse transcriptase-polymerase chain reaction analysis.

Measurements and Main Results: Gene expression analysis identified attributes not previously associated with lung development, such as chemokine-immunologic processes. Lung characteristics attributes (e.g., surfactant function) were observed at an earlier-than-anticipated age. We defined a 3,223 gene developing lung characteristic subtranscriptome capable of describing a majority of the process. In gene expression space, the samples formed a time-contiguous trajectory with transition points correlating with histological stages and suggesting the existence of novel molecular substages. Induction of surfactant gene expression characterized a pseudoglandular "molecular phase" transition. Individual gene expression patterns were independently validated. We predicted the age of independent human lung transcriptome profiles with a median absolute error of 5 days, supporting the validity of the data and modeling approach.

Conclusions: This study extends our knowledge of key gene expression patterns and bioontologic attributes underlying early human lung developmental processes. The data also suggest the existence of molecular phases of lung development.

Keywords: microarrays; surfactant; principal component analysis

In utero events, such as environmental exposures during critical windows of lung development, have been demonstrated to affect lung function and susceptibility to respiratory disease in later life (1–6). One of the widely proposed mechanisms behind this "fetal origins hypothesis" is alterations in gene expression in the developing fetus (7). To best understand the critical windows of early lung development and the molecular mecha-

AT A GLANCE COMMENTARY

Scientific Knowledge on the Subject

Knowledge of lung development in humans has been largely derived from gross anatomical staging with limited molecular studies. High throughput technologies have enabled comprehensive studies of organogenesis in animal models, but corresponding studies in humans are lacking.

What This Study Adds to the Field

We provide a unique data set significantly expanding current knowledge of gene expression patterns in the developing human lung in particular and in human organogenesis in general. From these data, we have identified essential attributes characterizing the biology of the lung, which should facilitate our understanding of lung disease pathogenesis.

nisms underlying these windows, we undertook a genome-wide study of gene expression in the normal fetal human lung.

The mechanisms regulating lung development are a temporally coordinated interaction of distinct functional modules (chemical and mechanical) that elicit the necessary morphological and physiological transformations from primordium to a differentiated, functional organ. Current knowledge of lung development has largely been derived from gross anatomical staging and limited molecular studies in humans, supplemented by careful histomorphologic, molecular, and genetic studies of animal models (8–11). The traditional foundation for describing mammalian lung development relies upon the delineation of this continuous process into distinct, morphologic stages based upon the presence of select histological features. Biological processes that transcend stages have long been appreciated. Several novel and canonical biochemical signaling pathways and morphogenic gradients previously associated with the developmental processes of other tissue-organ systems have been implicated in the lung (12–15).

Recently, high-throughput technologies (e.g., microarrays) have enabled comprehensive studies of genome-wide expression patterns and global biological processes underlying organogenesis of the mouse lung (16–18). Related studies of abnormal lung development in the mouse suggest the existence of essential regulatory modules that recur throughout development and transcend histomorphologic compartmentalization (19–22). However, there have been few systematic investigations of the global relationship between the gross morphology (macroscopic-scale features) and transcriptomic processes (microscopic-scale features) of the developing mammalian lung. Corresponding developmental studies in human are lacking.

In previous work (18), we used unsupervised principal component analysis (PCA) of genome-wide expression data to

(Received in original form July 14, 2009; accepted in final form October 7, 2009)

Supported by National Institutes of Health grants K25 HL91124, R01 HL88028, and P50 NS40828 (A.T.K.); K23 HG3983 (K.G.T.); R01 ES10855 (J.S.L.); U54 LM8748 (I.S.K.); U01 HL65899 (S.T.W.); and R01 HL71885 (T.J.M.).

* These authors contributed equally to this work.

Correspondence and requests for reprints should be addressed to Thomas J. Mariani, Ph.D., Dept. of Pediatrics, University of Rochester, 601 Elmwood Avenue, Box 703, Rochester, NY 14642. E-mail: tom_mariani@urmc.rochester.edu; or Scott T. Weiss, M.D., Channing Laboratory, Brigham and Women's Hospital, 181 Longwood Avenue, Boston, MA 02115.

This article has an online supplement, which is accessible in this issue's table of contents at www.atsjournals.org

Am J Respir Crit Care Med Vol 181, pp 54–63, 2010

Originally Published in Press as DOI: 10.1164/rccm.200907-1063OC on October 8, 2009

Internet address: www.atsjournals.org

identify global transcriptomic features of mouse lung development. Gene expression variation was found to be associated with macroscopic biological features such as age and the time-to-birth. Our analyses suggested the possibility of characterizing lung development in molecular, rather than histological, terms.

In this study we applied a similar approach to characterize genome-wide expression patterns in the developing human lung transcriptome using a set of fetal lung tissue specimens from National Institute of Child Health and Human Development-supported repositories. We assayed expression patterns from 38 unique samples, representing 29 distinct time-points, spanning the pseudoglandular and canalicular stages (estimated 53–154 days post conception [dpc]). Analysis of this developmental time interval was sufficient to objectively identify a minimal set of 3,223 characteristic genes, including anticipated and novel molecules, capable of defining essential features of human lung development. These characteristic genes were enriched for bioontologic attributes that are organ specific and traditionally associated with later lung development states. Moreover, we find evidence that molecular phases of human lung development parallel histologically defined stages but identify distinct developmental substages that were not previously appreciated.

METHODS

Human Tissue Samples and RNA Isolation

Human fetal lung tissues (Table 1) were obtained from two National Institute of Child Health and Human Development-supported tissue retrieval programs: the University of Washington Center for Birth Defects Research (Seattle, WA) and the University of Maryland Brain and Tissue Bank for Developmental Disorders (Baltimore, MD). Exclusion criteria include reported hepatic, renal, or pulmonary disease. The postmortem interval was less than 2 hours for samples from the University of Washington and less than 6 hours for samples from the University of Maryland. Creation of the tissue repository was approved by the University of Missouri–Kansas City Pediatric Institutional Review Board.

We studied 38 RNA samples generated from lung tissue obtained from 38 subjects of estimated gestational age between 53 and 154 days postconception (dpc). Estimated gestational age is based on the date of the last menstrual period as declared by the mother's self report at the time of consent. Between 20 and 30 mg of frozen tissue was homogenized, and total RNA was extracted with an on-column DNase I treatment according to the Qiagen RNeasy protocol (Qiagen, Valencia, CA) or the Illustra RNAspin Mini RNA Isolation Kit (GE Healthcare, Piscataway, NJ). The isolated RNA was run on an Agilent Bioanalyzer for quality and quantity assessment. RNA concentration was further assessed by measuring UV absorbance at a wavelength of 260 nm, and

TABLE 1. PHENODATA FOR TISSUE SAMPLES USED IN THIS STUDY

Gestation Age (d)	Source	Lab ID	Stage	Control = 0; Disease = 1/2	Array ID	Gender	Prediction	Postmortem Interval (h)
53	UW	H20029	P	0	992I		Train	<2
59	UW	H19128	P	0	944I	F	Train	<2
63	UW	H17657	P	0	707C		Test	<2
63	UW	H18338	P	0	707D	F	Train	<2
72	UW	H18419	P	0	707H	M	Train	<2
74	UW	H18877	P	0	944H	F	Test	<2
76	UW	H17470	P	0	707A	M	Test	<2
76	UW	H18401	P	0	707G		Train	<2
78	UW	H17574.2	P	0	944A	M	Train	<2
82	UW	H20022	P	0	992H		Train	<2
83	UW	H18606.2	P	0	944F	M	Test	<2
85	UW	H19908	P	0	944L	M	Train	<2
87	UW	H18480	P	0	944D		Test	<2
89	UW	H17920.2	P	0	944B	M	Train	<2
91	UW	H19967	P	0	992F	M	Test	<2
94	UW	H19998	P	0	992G		Train	<2
96	UW	H19313	P	0	944J	F	Train	<2
98	UW	H18875.3	P	0	944G	M	Train	<2
101	UW	H18402	P	0	992C	F	Test	<2
103	UW	H18356	P	0	707F	F	Test	<2
105	UW	H18450	P	0	707I	M	Train	<2
108	UW	H18355	P	0	707E		Test	<2
108	UW	H18475	P	0	707J	M	Test	<2
108	UW	H18475	P	0	992D	M	Train	<2
110	UW	H17486	P	0	707B	M	Train	<2
110	UW	H17486	P	0	992A	M	Test	<2
113	UW	H18484	P	0	944E	M	Train	<2
117	UW	H19120	C	0	992E	F	Train	<2
122	UW	H19896	C	0	944K	M	Test	<2
130	UM	UMB1621	C	0	738G	M	Train	1
133	UM	UMB1589	C	0	738F	F	Test	2
133	UM	UMB1631*	C	0	738H	M	Train	2
133	UM	UMB1756	C	0	823C	M	Test	1
133	UM	UMB1776*	C	0	823D	F	Train	1
134	UM	UMB1566	C	0	738E	M	Train	6
140	UM	UMB0045	C	0	738A	M	Test	2
140	UM	UMB1690	C	1 [†]	738I	M	Train	1
154	UM	UMB0684	C	2 [†]	738D	M	Test	5

Definition of abbreviations: C = canalicular; P = pseudoglandular; UM = University of Maryland; UW = University of Washington.

* Maternal smoking reported for UMB1631 and UMB1776.

[†] 1 = cause of death listed as anencephaly; 2 = cause of death listed as tuberous sclerosis.

RNA purity was assessed using the 260/280 ratio. Only RNA samples with an RNA concentration greater than 100 ng/μl and an RNA Integrity Number (RIN) greater than 6 were used for microarray analysis.

Microarray Profiling

The human lung transcriptome data are derived from whole lung tissue RNA of 38 unique human fetal samples spanning 29 distinct time points between 53 and 154 days dpc. RNA samples were profiled using the Affymetrix Human Genome GeneChip U133 Plus 2.0 microarray (Affymetrix, Santa Clara, CA) containing 54,675 probes corresponding to 19,501 unique NCBI Entrez genes. Labeled target was synthesized from purified RNA samples and hybridized according to the manufacturer's recommendations by the Harvard Medical School–Partners Healthcare Center for Genetics and Genomics using standard protocols (<http://www.hpcgg.org/Microarrays/resources>). Expression values were extracted from .CEL files using Robust Multi-array Average (RMA, for specific sample subsets) and Microarray Suite 5.0 (MAS 5.0) separately using BioConductor (<http://www.bioconductor.org>). The Affymetrix .CEL files and RMA normalized data matrix for these samples are accessible at the NCBI Gene Expression Omnibus repository (<http://www.ncbi.nlm.nih.gov/geo/>) (series accession number GSE14334).

The developing mouse lung transcriptome data of 24 sample profiles (six distinct time points, four replicate measurements each) has been published (16, 18) and is described in the online supplement.

PCA and Characteristic Gene Sets

We used PCA as a data reduction strategy for analysis of the data sets (see online supplement). The first principal component (PC1) defines the direction of greatest variation in the probe/gene transcriptomic feature space, with each subsequent PC explaining the next most variation. The contribution of a gene to a PC is proportional to its corresponding probe loading coefficient magnitude. PCA was applied on the standardized dataset without and with prior rank normalization separately, corresponding to the use of linear correlation and rank correlation as measures of similarity between samples, respectively. The resulting PCs are abbreviated reg (regular) PC and rank PC, respectively.

Gene Attribute Enrichment Analysis

We assessed gene sets for overrepresentation of specific bioontologic attributes using DAVID 2008 (<http://david.abcc.ncifcrf.gov>), data freeze October 2008 (23, 24). These attributes encompass gene ontologies, sequence motifs, protein domains, and biomolecular pathways. We consider an attribute to be significant if its Expression Analysis Systematic Explorer (EASE) score (Fisher exact test *P* value) is less than 0.05 relative to an appropriate background gene set.

Differential Expression Analysis of Array Data

Differential expression was defined in the full transcriptome of 54,675 probe sets using Significance Analysis of Microarrays (25) at a median FDR = 0 in TIGR Multi-experiment Viewer (<http://www.tm4.org/mev.html>) as we have previously described (18). Genes were further restricted to those demonstrating a fold change of 2.0 or greater in average expression.

Quantitative Reverse Transcriptase–Polymerase Chain Reaction

Quantitative reverse transcriptase–polymerase chain reaction (qPCR) was performed on a Stratagene MX3000P using Taqman chemistry as previously described (26). Inventoried (predeveloped) gene-specific assays for measuring gene expression were obtained from Applied Biosystems (Foster City, CA). Gene expression levels (dCT) were calculated relative to the measured CT value of PPIA (peptidyl prolyl isomerase A or cyclophilin A) as an internal, endogenous control.

RESULTS

Human Lung Development Transcriptome

The primary human lung development data set consists of microarray transcriptome profiles from whole lung tissue

RNA of 38 unique human fetal samples spanning 29 distinct time points, (53–154 dpc; Table 1). To minimize the effects of subject or age estimation–related variation in expression measurement and to model global gene-specific developmental expression patterns, we implemented probe-wise nonparametric regression using *locfit* (27) in BioConductor (see online supplement for details). This approach normalizes individual gene expression patterns using a global smoothing parameter function that minimizes a weighted least squares error. The resulting data, Dataset S1, represents individual smoothed trajectories for every transcript queried on the platform modeled at 28 distinct time points (identical to the time points of the primary dataset except 154 dpc) using all 38 arrays (see Figure E1 in the online supplement). The modeled expression profiles are available as supplemental data, and the preprocessed data set has been made publicly available (NCBI Gene Expression Omnibus repository series accession number GSE14334). These data represent a fetal human lung gene expression compendium.

Principal Components of the Developing Human Lung Transcriptome and Their Characteristic Genes

PCA was used to characterize the independent, dominant directions of maximal sample variation called principal components (PCs) in the developing lung transcriptome. PCA was applied to the dataset without and with prior rank normalization (18). The resulting PCs are abbreviated reg (regular) PC and rank PC. In each case, PC1–3 accounted for a substantial amount of total sample variation (75.39% for reg PC and 72.29% for rank PC). The PC1 sample coordinates were significantly correlated (*P* < 0.001) with fetal age (correlations of 0.9874 for reg PC1 and 0.9633 for rank PC1). This is anticipated for a data set with the major underlying feature of experimental design being subject age and supports the validity of the smoothing and PCA modeling approach.

We sought to identify the key genes contributing to expression variation over development. The contribution of a gene to a PC is proportional to the gene's corresponding microarray probe loading coefficient (Dataset S3). For each PC, we defined the top 5% (ranks 1–975) of genes with the highest loading coefficient magnitudes as its characteristic genes. Genes contributing substantially to PC1 included surfactant proteins SFTPB and SFTPC, a member of the Wnt-binding Frizzled-related proteins SFRP2, the catenin/plakophilin CTNND2, and the scavenger receptor CD36. Genes contributing substantially to PC2 included the surfactant-associated claudin CLDN18, the chemokine CXCL5, and the MHC class II molecules HLA-DRA and HLA-DRB1. Genes contributing substantially to PC3 included the general transcription factor FOSB. We observed a small set of common genes (22.5–30.5%) among the characteristic genes of PC1–3 (Table E1). These may represent a set of molecules contributing to multiple, overlapping processes important to early lung development. Forty-six genes were common to all PCs, including prominently ranked surfactant-related proteins (SFTPB, SFTPC, and CLDN18), immune molecules, and a number of molecular transporters (Table 2).

Bioontologic Attributes of the PCs of the Developing Human Lung Transcriptome

It is difficult, and potentially misleading, to resolve complex biological processes described by these PCs using individual genes. Therefore, we identified biological correlates of PC1–3 using gene attribute enrichment analysis. These biological correlates can be macroscopic (e.g., development) or micro-

TABLE 2. GENES COMMON TO ALL CHARACTERISTIC GENE SETS OF REGULAR AND RANK PRINCIPAL COMPONENTS 1–3 OF DATASET S1 (INCLUDING ENTREZ ID, GENE SYMBOL, AND RANK IN EACH PRINCIPAL COMPONENT)

ID	Gene	Reg PC1 40.4%	Reg PC2 25.6%	Reg PC3 9.4%	Rank PC1 35.5%	Rank PC2 27.5%	Rank PC3 9.3%
213	ALB	98	6	13	3	4	1
336	APOA2*	599	408	179	55	141	3
361	AQP4*	194	3	29	690	10	131
443	ASPA	122	210	133	446	389	576
604	BCL6	345	833	197	246	745	815
894	CCND2	240	86	72	415	100	120
960	CD44	56	709	545	203	706	401
993	CDC25A	163	64	381	117	8	95
1052	CEBPD	146	414	451	126	21	617
1501	CTNND2	12	129	89	11	7	33
2348	FOLR1	154	361	55	733	866	219
3113	HLA-DPA1†	65	37	177	402	81	178
3122	HLA-DRA†	52	8	54	193	23	119
3123	HLA-DRB1†	95	22	45	422	98	180
3762	KCNJ5*	40	353	129	81	147	137
3772	KCNJ15*	8	13	86	15	52	257
4170	MCL1*	818	76	128	332	231	70
6374	CXCL5	51	420	65	20	17	566
6423	SFRP2	10	41	3	53	124	22
6439	SFTPB‡	2	7	4	144	194	74
6440	SFTPC‡	1	1	7	48	61	90
8417	STX7*	247	165	288	221	817	184
8707	B3GALT2	703	88	305	642	45	363
9022	CLIC3*	31	9	6	138	46	19
9123	SLC16A3*	544	936	216	474	144	96
9645	MICAL2	459	268	88	580	311	299
10048	RANBP9	957	162	971	509	260	538
10396	ATP8A1*	72	59	345	151	347	718
10568	SLC34A2*	19	203	25	427	560	691
22936	ELL2	632	517	338	143	265	67
23250	ATP11A*	644	154	27	765	498	123
25890	ABI3BP	27	285	268	13	471	904
27074	LAMP3	7	23	26	195	210	313
51208	CLDN18‡	37	29	42	96	67	97
51512	GTSE1	262	711	402	257	867	268
53405	CLIC5*	45	757	505	100	343	312
54959	ODAM	923	206	357	583	28	159
55635	DEPDC1	92	890	382	76	271	367
56884	FSTL5	183	665	116	30	20	50
64073	C19orf33	74	496	775	161	217	483
79689	STEAP4	4	265	242	7	15	176
84663	CYorf15B	289	67	21	62	489	177
84803	LPAAT-τ	104	331	189	45	24	62
84879	MFSD2	255	211	144	653	305	416
259266	ASPM	251	951	48	512	749	7
493869	LOC493869	9	2	57	10	1	13

Definition of abbreviations: PC = principal component; Reg = regular.

* Substrate-specific transporter activity (n = 12).

† Major histocompatibility complex (MHC) class II complex (n = 3).

‡ Surfactant associated (n = 3).

scopic (e.g., nucleic acid binding) in scale. For instance, we noted above that the PC1 coordinates of the samples were correlated with fetal age, a macroscopic variable.

Gene attribute enrichment analysis was performed on each characteristic gene set of reg and rank PC1–3, relative to the background gene set of all 19,501 unique Entrez genes measured on the U133 Plus 2.0 microarray. A total of 1,084 unique attributes form the union of all six characteristic gene sets (Table E2), while 35 attributes were common to all six PCs (Tables 2 and E3). Among the common characteristic attributes, we noted the abundance of surfactant–gas exchange attributes, even though the function of surfactant proteins in these early stages of lung development is unclear. Also prominent are immunologic-major histocompatibility complex (MHC), in particular MHC class II genes, whose role in early

lung development is similarly unclear. Regarding attributes that are unique to individual PCs, PC1 was enriched for DNA and nuclear organization and replication and cell cycle processes. PC2 was enriched for cytoskeletal protein domains (e.g., spectrins) and transcription factor activity and binding motifs. PC3 was enriched for cell-specific differentiation processes, cellular motion, and extracellular matrix interaction (e.g., chondroitin sulfate proteoglycans) attributes.

We defined a developing lung characteristic subtranscriptome (DLCS, Table E4) as the union of characteristic genes of PC1–3. This set is comprised of 3,223 unique genes, corresponding to 11,558 probes. The DLCS constitutes 16.53% of the total 19,501-gene transcriptome measured by the microarray, and its corresponding 11,588 probes account for 72.73% (reg PC) or 69.31% (rank PC) of the variation of PC1–3 on average.

Figure 1. Global gene expression patterns in fetal human lung development. (A) Results for unsupervised hierarchical clustering of gene expression trajectories for 403 representative genes with a high-loading coefficient (top 5%) in at least one principal component (PC) and fold-change greater than four across the time-series. Rows represent individual genes clustered by expression. Columns represent individual time-points ordered by age from 52 to 140 days post conception (dpc). Black arrow indicates the transition from the pseudoglandular (Pseudo) to canalicular (Canal) stages of lung development. Normalized gene expression is indicated by the key (bottom), with red representing high expression levels and green representing low expression levels. (B) High magnification views from (A) highlighting genes with expression significantly changing across the molecular phase transition from early to late pseudoglandular stage (gray arrow). The top image (blue box) is one cluster of 17 representative genes decreasing during the pseudoglandular stage. The bottom image (orange box) is one cluster of 39 representative genes increasing during the pseudoglandular stage.

Temporal expression patterns for a subset of these genes are presented in Figure 1A. Many of these DLCS genes showed marked changes in expression at the transition from the pseudoglandular to canalicular stage of lung development.

Discovery of Transcriptomic or Molecular Phases of Human Lung Development

Each array can be viewed as a point in space defined by genome-wide variation in expression (e.g., as summarized by PCA), demonstrating the relationship between samples and ages. When the developing human lung samples are projected in PCA gene expression space, there is a strong and significant correlation between PC1 and age. However, the configuration of the samples plotted in principal component space is highly complex in structure (Figures 2 and E2). One prominent feature of this visualization is the identification of two periods where progression along the PC1 (e.g., age) axis “slows” dramatically, resulting in a transition point of the lung development trajectory. The first of these is between Days 91 and 96, and the second is between Days 110 and 117. The second transition point coincides with the transition from the pseudoglandular to canalicular stages of lung development (16 weeks or 112 days). The first transition point occurs within the middle of the pseudoglandular stage, suggesting a previously unappreciated molecular subphase of early human lung development.

To confirm this trajectory is not stochastic, PCA was performed as above but was restricted to the DLCS feature set. PC1–3 here accounted for 80.41% (reg PC) or 79.04% (rank PC) of total sample variation. The DLCS PC1 sample coordinates were significantly ($P < 0.001$) correlated with fetal age (correlation 0.9956 for reg PC1 and 0.9956 for rank PC1). The resulting PC1–3 sample configurations were highly similar to those observed using the full transcriptome feature set (Figures 2B and E3). These observations suggest that a small and specific subset of the transcriptome—the DLCS—recapitulates the essential macroscopic features of human lung development from 53 to 140 dpc. Consistent with the hypothesis that two distinct molecular phases of the pseudoglandular stage exist, recreating the time-contiguous trajectory of the samples using the DLCS alone, a set enriched in informative genes (and presumably with reduced noise), identifies the identical transition points.

We identified genes with significantly different expression between these putative molecular phases of the pseudoglandular stage of lung development. To avoid confounding this analysis by including transitional time points, we defined the early pseudoglandular molecular phase as between 63 and 85 days and the late pseudoglandular molecular phase as between 101 and 108 days. We identified 245 genes with significantly higher expression from Days 63 to 85 and 230 genes with significantly higher expression from Days 101 to

Figure 2. Principal component analysis of human lung development in gene expression space. Time points from Dataset S1 are presented in the principal component (PC)1-PC2 plane defined using (A) the full transcriptome (19,501 unique genes) or (B) the developing lung characteristic gene set (DLCS) (3,223 unique genes). Lines are drawn between consecutive time points, representing a time-contiguous trajectory of the developmental process. Axes show the percentage sample variance captured by each PC when the full transcriptome or developing lung characteristic subtranscriptome (DLCS) is used for regular or rank principal component analysis.

108 (Figure 1B and Table E5). Attribute analyses of these genes demonstrated that the early pseudoglandular genes were enriched with chromosomal organization processes associated with mitosis, and the late pseudoglandular genes were enriched with surfactant function–gas exchange and immunological-MHC class II attributes (Table E6). All surfactant protein genes measured (SFTPA, B, C, and, D) were significantly increased from the early to late pseudoglandular phase. Together, these data suggest the existence of two distinct molecular phases within the pseudoglandular stage of human lung development, which can be distinguished by induction of surfactant gene expression.

Validation of DLCS Gene Expression Profiles in the Developing Embryonic Human Lung

qPCR was used to independently verify the expression profile of 18 DLCS genes in the developing human lung at 28 distinct time points. Genes were selected for their prominent ranking in the characteristic gene sets of PC1–3 or for their corresponding, significantly overrepresented bioontologic attributes, mainly surfactants and immunologic/stress response. To test the accuracy of the array-based global expression profiles, we computed the correlation of microarray-reported signal and negative qPCR delta CT (dCT) values across these 26 time points (Figure 3 and Table E7). Fifteen of 18 genes showed significantly correlated expression profiles: CCL20 ($r = 0.572$; $P = 0.002$), CD36 ($r = 0.863$; $P < 0.001$), CFTR ($r = 0.519$; $P = 0.007$), CLDN18 ($r = 0.6985$; $P = 0.001$), CLIC5 ($r = 0.430$; $P = 0.029$), CTNND2 ($r = 0.757$; $P < 0.001$), CXCL3 ($r = 0.6527$; $P < 0.001$), CXCL5 ($r = 0.5795$; $P = 0.001$), CXCL6 ($r = 0.533$; $P = 0.005$), HPGD ($r = 0.501$; $P = 0.009$), KITLG ($r = 0.619$; $P < 0.001$), SERPINA1 ($r = 0.511$; $P = 0.007$), SFTPB ($r = 0.7880$; $P < 0.001$), SFTPC ($r = 0.8927$; $P < 0.001$), and TUBB2B ($r = 0.8051$; $P < 0.001$).

We also used qPCR to test the validity of the novel pseudoglandular substage, or molecular phase, and the ability of individual genes to define this phase transition. We evaluated 10 genes predicted to be differentially expressed across the substages (Figure 3 and Table E8), confirming that seven have significantly different expression levels: CCL20 ($P = 0.024$), CD36 ($P < 0.001$), CLDN18 ($P = 0.011$), CTNND2 ($P = 0.002$), SFTPB ($P = 0.035$), SFTPC ($P = 0.005$), and TUBB2B ($P < 0.001$).

Estimating the Age/Developmental Maturity of the Human/Mouse Lung by Transcriptome Profiling

We investigated the robustness of this characterization of early lung development by assessing its capacity to (1) estimate the

age of independent human fetal lung samples and (2) chronologically order developing mouse lung samples on the basis of their lung transcriptome profile alone. To do this, we performed a holdout cross validation, whereby the developing human lung data were split into training (22 arrays) and test (16 arrays) sets. Here, the training set (only) was used to regenerate smoothed time-dependent, whole-genome expression trajectories using probewise, nonparametric regression (as for Dataset S1). Dataset S2 represents individual smoothed trajectories for every transcript queried on the platform modeled at 18 distinct time points, at regular 5-day intervals from 59 to 140 dpc, using the 22 training arrays (Figure E1). Dataset S2 was used to rederive the transcriptomic PC1–3 feature space of lung development, and the human or mouse lung test sample transcriptome profiles were projected into this space as we have previously described (28, 29).

To estimate the age of human samples, PCA was performed on the training set full transcriptome (19,501 genes, 54,675 probes) or the training set limited to the DLCS (3,223 genes, 11,558 probes). Next, all 38 human fetal lung samples were projected into the transcriptomic PC1–3 space of the training set (Figures E4 and E5). The estimated age of all human samples was inferred from the age of its closest training sample in PC1–3 space (Figure 4A). The median absolute differences between estimated and true age of the 22 training samples were 5 days for the full transcriptome and the DLCS. For the 16 primary test samples, the median absolute differences were 7 days (full transcriptome) and 5 days (DLCS), respectively. These results demonstrate the ability of the PCA-based time-dependent trajectory to accurately define human fetal lung age and confirm the added utility of the DLCS genes to describe this biological process.

To assess the age of mouse lung samples, PCA was performed on the training set (Dataset S2) limited to the set of 7,683 orthologous genes (in common between the human and mouse microarray platforms) or those orthologs in the DLCS (1,568 genes). The test data for this purpose consisted of 21 developing mouse whole lung transcriptome profiles at six discrete time points: embryonic Days 14 (×4), 17 (×3), and 19 (×4) and postnatal Days 7 (×3), 14 (×3), and 28 (×4) (16, 18). When projected in human lung development transcriptomic PC space, the mouse samples, especially the embryonic time points (14, 17, and 19 dpc), were correlated with their true developmental age (Figures 4B, E6, and E7). The rank correlation of PC1 coordinates of mouse samples (embryonic and postnatal, median per time point) with their true age is 0.9429 ($P < 0.01$). These results further validate the biological foundation of this PCA-based, time-contiguous trajectory and

Figure 3. Human lung development expression profiles for select developing lung characteristic gene set (DLCS) genes, including those with pseudoglandular substage specific changes. Individual gene expression profiles (modeled and median normalized) for (A) surfactant-related, (B) chemokine, and (C) substage specific genes. (D) Quantitative data on qPCR-based validation for individual gene expression profiles. For a given gene, the RMA-derived signal intensity was correlated (Corr; r) with the gene expression level determined by quantitative reverse transcriptase–polymerase chain reaction.

suggest that it captures conserved biological features of lung development.

DISCUSSION

We present here, for the first time, a comprehensive assessment of gene expression patterns in the developing human lung between 53 and 154 dpc, a time interval spanning the pseudoglandular and canalicular stages. These data, including thousands of individual gene expression patterns, are publicly available. We have applied novel nonparametric, regression-based smoothing to the primary dataset in an effort to minimize subject/sample and age estimation-dependent effects. This approach effectively improves dataset quality as objectively defined by increasing similarity between trajectories for multiple probe sets interrogating the same transcript (data not shown). The resulting dataset significantly expands our current understanding of changes in the expression of genes during human lung development.

We distilled the developing human lung transcriptome into independent, dominant directions of gene expression variation using unsupervised PCA. Analyzing the developmental time series in this way, we can objectively resolve the contribution of

different biological/functional modules throughout the course of development. For each PC, a characteristic gene set was defined and its corresponding bioontologic attribute profile identified. The characteristic gene sets were significantly enriched for many attributes known to be involved in lung development, but whose roles in these early stages have not been well characterized (e.g., surfactant). In addition, a number of genes previously unappreciated as playing a role in lung development were discovered using this approach.

Our PCA characterization facilitated visualization of the time-series as a linear trajectory defined by gene expression variation. The most dominant direction of sample variation in this developing human lung transcriptome (PC1) was strongly correlated with estimated gestational age. This is intuitive and consistent with our earlier studies of the developing mouse lung transcriptome time series (18). However, the short gestational period of the mouse coupled with the low sampling frequency of that dataset limited the ability to detect finer transcriptomic granularity in that process.

Traditional embryology partitions growth of the developing human lung into five stages: embryonic (26 dpc to 5 weeks), pseudoglandular (5–16 weeks), canalicular (16–26 weeks), sacular (26 weeks to birth), and alveolar (birth to 6 months), based upon selective morphological features that are observed

Figure 4. Principal component analysis-based lung development sample age prediction. The rank principal component (PC)1–3 transcriptomic space was defined using modeled data from 22 human fetal lung training samples (Dataset S2). (A) The 16 human lung samples defined as the test set were projected into the rank PC1–3 transcriptomic space of Dataset S2. The Euclidian distance in rank PC1–3 between each test sample (rows) and all training samples (columns) was determined and is displayed; green represents a shorter distance, and black represents a farther distance. The training set time point closest to each test sample was defined to be its estimated age. This analysis was performed using the full transcriptome (Tx) and the developing lung characteristic gene set (DLCS) separately. The median absolute difference between estimated and actual age of the 16 test samples are 7 (for full Tx) and 5 (for DLCS) days, respectively. (B) Using the 1,568-gene DLCS features common to human and mouse array data sets, 21 mouse lung samples were projected into the rank PC1–3 space of Dataset S2. As in (A), the distance between Dataset S2 time points and developing mouse lung transcriptome profiles at 14, 17, and 19 days post conception (dpc)(e14–19) and postnatal Days 7, 14, 28 (p7–28) was determined and is displayed, where red represents a shorter distance and black represents a farther distance.

within these intervals. However, it has been long appreciated that complex biological processes occurring during lung development are not specifically associated with this classical stage definition. In the human data, we observed a gene expression transition point at approximately 117 dpc, reflecting the age associated with the morphologic transition from the pseudoglandular to canalicular stages. A second transition point was noted at approximately 94 dpc, suggesting the presence of at least two distinct molecular phases within the pseudoglandular stage. The substages identified by this PCA characterization might reflect critical molecular windows of lung development (3) and suggest that molecular staging in this manner captures information that is complementary but distinct from classic morphologic staging alone. Given sufficient sampling/measurement frequency, PCs based upon genome-wide expression data could form a rational basis for an alternative and objective molecular taxonomy of lung development.

To further explore the biology identified by our analysis of the developing human lung transcriptome, we focused on characteristic gene sets (the highest 5% loading coefficient magnitude genes) of PC1–3 and their ontological attributes. We defined the 3,223–gene union of PC1–3 characteristic genes as the developing lung characteristic subtranscriptome (DLCS), which may be regarded as the minimal set of genes that accounts for a significant proportion of the transcriptomic variation in the early fetal developing human lung. Among the 46 genes common to all PC1–3 characteristic gene sets are three surfactant-associated genes (STFPB, SFTPC, and CLDN18), three MHC class II genes, and 12 substrate-specific transporter genes. Overall, the DLCS contained 28 of 77 genes previously identified to be functionally involved in general lung development (14), which is a 2.9-fold enrichment (odds ratio $P < 0.5 \times 10^{-4}$ by two-tailed Fisher exact test; Table E9). This enrichment for functionally relevant genes and the prominence of surfactants strongly suggests the PC characterization to be

a meaningful abstraction of the specific molecular biology underlying lung development.

There were 35 bioontologic attributes common to all PC1–3 characteristic gene sets. Not surprising among these were attributes related to cell cycle and cellular division, universal to general developmental processes. Additionally, there was a significant presence of surfactant–gaseous exchange and immunologic–MHC class II attributes. Genes with these attributes possess expression profiles that were largely increasing from 53 to 140 dpc. Although the production of surfactant is associated with later gestational development, our observations that these surfactant proteins are expressed relatively early in fetal lung development is consistent with a prior description of SFTPB and SFTPC expression as early as Week 13 in humans (30) and before the end of the pseudoglandular stage in mice (31). The early expression of surfactant proteins may also indicate a necessity of early molecular programming of the lung for subsequent development. This is supported by recent work revealing the potential of embryonic stem cells to form glandular respiratory epithelium after coculture with dissociated fetal lung (32). Additionally, given the recent description of forkhead box M1 (FOXM1) as a key regulator of surfactant protein expression (33), we found that FOXM1 is a PC1 characteristic gene here, further supporting the role of surfactants in early lung development.

We addressed the robustness of our findings in multiple ways. First, we noted that PCA using the DLCS, a feature set enriched in informative expression profiles likely to increase the signal-to-noise ratio, returned a time-contiguous trajectory very similar to that of the full transcriptome. qPCR for DLCS genes demonstrated a very high rate of global validation for individual gene expression trajectories (83%) and for differential expression of substage marker genes (70%). Next, we assessed the validity of this PCA-based characterization by testing its ability to predict the age of independent developing human or mouse

TABLE 3. SIGNIFICANT BIOONTOLOGIC ATTRIBUTES COMMON TO ALL CHARACTERISTIC GENE SETS OF REGULAR AND RANK PRINCIPAL COMPONENTS 1–3 OF DATASET S1*

Category - Term [†]	Reg PC1 40.4%	Reg PC2 25.6%	Reg PC3 9.4%	Rank PC1 35.5%	Rank PC2 27.5%	Rank PC3 9.3%
COG_ONTOLOGY - cell division and chromosome partitioning	1.9	1.8	2.0	1.9	1.9	2.3
GOTERM_BP - GO:000087~M phase of mitotic cell cycle	5.2	1.7	2.0	3.5	1.8	1.6
GOTERM_BP - GO:0000278~mitotic cell cycle	4.4	1.8	1.8	3.0	1.8	1.5
GOTERM_BP - GO:0000279~M phase	4.7	1.5	1.8	3.1	1.6	1.7
GOTERM_BP - GO:0002504~antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	4.8	4.9	5.1	7.6	7.6	7.8
GOTERM_BP - GO:0007049~cell cycle	2.9	1.5	1.6	2.0	1.4	1.4
GOTERM_BP - GO:0008283~cell proliferation	2.0	1.7	1.3	1.6	1.6	1.3
GOTERM_BP - GO:0016043~cellular component organization and biogenesis	1.2	1.3	1.3	1.1	1.2	1.2
GOTERM_BP - GO:0022402~cell cycle process	2.9	1.5	1.5	2.1	1.4	1.3
GOTERM_BP - GO:0022403~cell cycle phase	4.3	1.6	1.9	3.0	1.7	1.6
GOTERM_BP - GO:0050789~regulation of biological process	1.1	1.2	1.2	1.1	1.1	1.2
GOTERM_BP - GO:0050828~regulation of liquid surface tension	16.9	17.1	13.5	13.4	17.8	13.6
GOTERM_BP - GO:0051301~cell division	4.7	2.0	2.1	3.0	1.8	1.8
GOTERM_CC - GO:0005856~cytoskeleton	1.6	1.4	1.3	1.3	1.4	1.4
GOTERM_CC - GO:0042613~MHC class II protein complex	6.3	6.2	6.7	9.8	9.6	9.9
GOTERM_CC - GO:0043228~nonmembrane-bound organelle	1.7	1.2	1.3	1.4	1.3	1.3
GOTERM_CC - GO:0043232~intracellular nonmembrane-bound organelle	1.7	1.2	1.3	1.4	1.3	1.3
GOTERM_MF - GO:0005515~protein binding	1.2	1.3	1.2	1.1	1.2	1.2
GOTERM_MF - GO:0032395~MHC class II receptor activity	8.9	8.9	7.4	10.9	10.7	11.1
INTERPRO - IPR001003:MHC class II, alpha chain, N-terminal	10.9	10.6	11.2	14.5	14.6	14.9
INTERPRO - IPR002946:Intracellular chloride channel	9.1	8.8	9.4	9.1	12.2	9.3
INTERPRO - IPR014745:MHC class II, alpha/beta chain, N-terminal	7.2	7.1	7.5	10.9	11.0	11.2
PIRSUPERFAMILY - PIRSF001991:class II histocompatibility antigen	5.7	6.4	6.2	9.5	10.3	9.5
SP_PIR_KEYWORDS - alternative splicing	1.1	1.3	1.2	1.2	1.3	1.2
SP_PIR_KEYWORDS - cell cycle	4.0	1.6	2.1	2.5	1.7	1.6
SP_PIR_KEYWORDS - cell division	5.5	2.3	2.3	3.4	2.2	1.8
SP_PIR_KEYWORDS - chromosomal rearrangement	1.8	2.8	1.9	1.9	2.4	2.0
SP_PIR_KEYWORDS - gaseous exchange	17.5	17.7	13.9	13.7	18.0	13.9
SP_PIR_KEYWORDS - heterodimer	2.5	2.9	3.2	2.4	2.5	2.4
SP_PIR_KEYWORDS - mhc ii	6.4	6.4	6.7	10.0	9.8	10.1
SP_PIR_KEYWORDS - phosphoprotein	1.4	1.5	1.2	1.3	1.3	1.3
SP_PIR_KEYWORDS - pulmonary surfactant	17.5	17.7	13.9	13.7	18.0	13.9
SP_PIR_KEYWORDS - surface film	17.5	17.7	13.9	13.7	18.0	13.9
SP_PIR_KEYWORDS - ubl conjugation	2.7	1.8	2.2	2.4	1.7	2.1
UP_SEQ_FEATURE - mutagenesis site	1.4	1.4	1.3	1.3	1.3	1.3

See Table 2 for definition of abbreviations.

* Their fold enrichment (odds ratio of 2×2 confusion matrix for Fisher exact test of overrepresentation) in each PC is indicated. This is not a gene attribute enrichment analysis of the 46 genes from Table 2.

[†] Major histocompatibility complex (MHC) class II and surfactant associated attributes are in boldface type.

lung samples. We divided the 38 fetal human lung samples into training (22 samples) and test (16 samples) sets and used the principal components of the training set to predict gestational age in the test set. We found that we could accurately estimate the gestational age of the test sample using the 3,223 gene DLCS or the full (genome-wide) transcriptome. The ability to independently predict gestational age from a lung transcriptomic profile may be viewed as proof of the concept that the PC characterization captures the essential biology of the system because one expects age to be critical parameter in the lung development process. The ability to chronologically order developing mouse lung profiles by their age in the human lung development PC space further suggests the DLCS identifies conserved molecular mechanisms and bioontologic attributes.

There are several limitations to this study. First, the dataset is limited to a relatively small number of samples from a specific window of time. We analyzed only 38 samples, including 29 distinct time points, over a 101-day sampling age range solely from early fetal stages. A larger sampling window and a higher sampling frequency would add to the details of the analysis and may lend further support for the presence of additional

molecular stages that we are unable to investigate more comprehensively with the current data. However, the lack of available human samples during late fetal gestation means that extending the embryonic time points beyond the current age window is essentially impossible. We have performed preliminary microarray analysis of newborn and juvenile human lung tissue. Data from these samples is confounded by multiple variables, including varying age at birth and primary or secondary pulmonary complications. Preliminary analysis of data from these samples suggests they represent a poor comparison group for the fetal data set, and thus we have decided against including them in the current study. Second, there may be a degree of error in the estimation of fetal gestational age. This and other factors, such as maternal smoking history (34), may affect gene expression in individual samples. However, our findings were robust in the prediction of gestational age. In fact, we implemented an expression profile smoothing approach specifically to minimize variability due to sampling. Third, we did not dissect individual cellular contributors to fetal lung development. Instead, we defined the process as the dynamic sum of all cellular constituents. Whole lung tissue transcriptome profiles may

have limited success in resolving cell/tissue specific molecular events, particularly during the later stages of development when the lung cell population is more heterogeneous with distinct cooccurring and specialized molecular processes. Nonetheless, our data are representative of the overall molecular events occurring in the early fetal human lung.

In conclusion, we present novel expression patterns for thousands of genes in the developing normal human lung. Our analyses of the human fetal lung transcriptome support current developmental paradigms and have revealed the existence of molecular phases of development. Our results have been validated via different modalities and appear robust. These observations, and the application of these analytic strategies to more comprehensive data sets, will further our understanding of essential molecular events in normal development and pathological derangements of the lung and will provide further insights into critical windows of development wherein environmental exposures may lead to subsequent disease.

Conflict of Interest Statement: None of the authors has a financial relationship with a commercial entity that has an interest in the subject of this manuscript.

Acknowledgment: The authors thank Alyssa McCoy for technical assistance and B. Raby and R. Lazarus for analytical advice.

References

1. Dietert RR, Etzel RA, Chen D, Halonen M, Holladay SD, Jarabek AM, Landreth K, Peden DB, Pinkerton K, Smialowicz RJ, *et al.* Workshop to identify critical windows of exposure for children's health: immune and respiratory systems work group summary. *Environ Health Perspect* 2000;108:483–490.
2. Maritz GS, Morley CJ, Harding R. Early developmental origins of impaired lung structure and function. *Early Hum Dev* 2005;81:763–771.
3. Pinkerton KE, Joad JP. The mammalian respiratory system and critical windows of exposure for children's health. *Environ Health Perspect* 2000;108:457–462.
4. Shi W, Chen H, Sun J, Chen C, Zhao J, Wang YL, Anderson KD, Warburton D. Overexpression of Smurf1 negatively regulates mouse embryonic lung branching morphogenesis by specifically reducing Smad1 and Smad5 proteins. *Am J Physiol Lung Cell Mol Physiol* 2004;286:L293–L300.
5. Stick S. Pediatric origins of adult lung disease. 1. The contribution of airway development to paediatric and adult lung disease. *Thorax* 2000;55:587–594.
6. Torday JS, Rehan VK. Developmental cell/molecular biologic approach to the etiology and treatment of bronchopulmonary dysplasia. *Pediatr Res* 2007;62:2–7.
7. Hall JG. The importance of the fetal origins of adult disease for geneticists. *Clin Genet* 2007;72:67–73.
8. Cardoso WV, Lu J. Regulation of early lung morphogenesis: questions, facts and controversies. *Development* 2006;133:1611–1624.
9. Massaro GD, Massaro D. Formation of pulmonary alveoli and gas-exchange surface area: quantitation and regulation. *Annu Rev Physiol* 1996;58:73–92.
10. Warburton D, Bellusci S, De Langhe S, Del Moral PM, Fleury V, Mailleux A, Tefft D, Unbekandt M, Wang K, Shi W. Molecular mechanisms of early lung specification and branching morphogenesis. *Pediatr Res* 2005;57:26R–37R.
11. Perl AK, Whitsett JA. Molecular mechanisms controlling lung morphogenesis. *Clin Genet* 1999;56:14–27.
12. Cardoso WV. Molecular regulation of lung development. *Annu Rev Physiol* 2001;63:471–494.
13. Horowitz A, Simons M. Branching morphogenesis. *Circ Res* 2008;103:784–795.
14. Maeda Y, Dave V, Whitsett JA. Transcriptional control of lung morphogenesis. *Physiol Rev* 2007;87:219–244.
15. Miura T. Modeling lung branching morphogenesis. *Curr Top Dev Biol* 2008;81:291–310.
16. Bonner AE, Lemon WJ, You M. Gene expression signatures identify novel regulatory pathways during murine lung development: implications for lung tumorigenesis. *J Med Genet* 2003;40:408–417.
17. 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;26:541–548.
18. Kho AT, Bhattacharya S, Mecham BH, Hong J, Kohane IS, Mariani TJ. Expression profiles of the mouse lung identify a molecular signature of time-to-birth. *Am J Respir Cell Mol Biol* 2009;40:47–57.
19. Dave V, Childs T, Xu Y, Ikegami M, Besnard V, Maeda Y, Wert SE, Neilson JR, Crabtree GR, Whitsett JA. Calcineurin/Nfat signaling is required for perinatal lung maturation and function. *J Clin Invest* 2006;116:2597–2609.
20. Wan H, Luo F, Wert SE, Zhang L, Xu Y, Ikegami M, Maeda Y, Bell SM, Whitsett JA. Kruppel-like factor 5 is required for perinatal lung morphogenesis and function. *Development* 2008;135:2563–2572.
21. Kourou-Mehr H, Werb Z. Candidate regulators of mammary branching morphogenesis identified by genome-wide transcript analysis. *Dev Dyn* 2006;235:3404–3412.
22. Otulakowski G, Duan W, O'Brodovich H. Global and gene-specific translational regulation in rat lung development. *Am J Respir Cell Mol Biol* 2009;40:555–567.
23. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol* 2003;4:3.
24. Hosack DA, Dennis G Jr, Sherman BT, Lane HC, Lempicki RA. Identifying biological themes within lists of genes with EASE. *Genome Biol* 2003;4:R70.
25. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 2001;98:5116–5121.
26. Simon DM, Arkan MC, Srisuma S, Bhattacharya S, Andalicio T, Shapiro SD, Mariani TJ. Epithelial cell PPARgamma is an endogenous regulator of normal lung maturation and maintenance. *Proc Am Thorac Soc* 2006;3:510–511.
27. Loader C. 1999. Local regression and likelihood, statistics and computing. New York: Springer.
28. Kho AT, Zhao Q, Cai Z, Butte AJ, Kim JY, Pomeroy SL, Rowitch DH, Kohane IS. Conserved mechanisms across development and tumorigenesis revealed by a mouse development perspective of human cancers. *Genes Dev* 2004;18:629–640.
29. Liu H, Kho AT, Kohane IS, Sun Y. Predicting survival within the lung cancer histopathological hierarchy using a multi-scale genomic model of development. *PLoS Med* 2006;3:e232.
30. Liley HG, White RT, Warr RG, Benson BJ, Hawgood S, Ballard PL. Regulation of messenger RNAs for the hydrophobic surfactant proteins in human lung. *J Clin Invest* 1989;83:1191–1197.
31. Perl AK, Wert SE, Nagy A, Lobe CG, Whitsett JA. Early restriction of peripheral and proximal cell lineages during formation of the lung. *Proc Natl Acad Sci USA* 2002;99:10482–10487.
32. Denham M, Cole TJ, Mollard R. Embryonic stem cells form glandular structures and express surfactant protein C following culture with dissociated fetal respiratory tissue. *Am J Physiol Lung Cell Mol Physiol* 2006;290:L1210–L1215.
33. Kalin TV, Wang IC, Meliton L, Zhang Y, Wert SE, Ren X, Snyder J, Bell SM, Graf L Jr, Whitsett JA, *et al.* Forkhead Box m1 transcription factor is required for perinatal lung function. *Proc Natl Acad Sci USA* 2008;105:19330–19335.
34. Breton CV, Byun H-M, Wenten M, Pan F, Yang A, Gilliland FD. Prenatal tobacco smoke exposure affects global and gene-specific DNA methylation. *Am J Respir Crit Care Med* 2009;180:462–467.