Expression Profiles of the Mouse Lung Identify a Molecular Signature of Time-to-Birth

Alvin T. Kho^{1*}, Soumyaroop Bhattacharya^{2*}, Brigham H. Mecham², Jungha Hong², Isaac S. Kohane¹, and Thomas J. Mariani²

¹Children's Hospital Informatics Program, Children's Hospital Boston and Harvard-MIT Division of Health Sciences and Technology, and

²Pulmonary Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts

A greater understanding of the regulatory processes contributing to lung development could help ameliorate morbidity and mortality in premature infants and identify individuals at risk for congenital and/or chronic lung diseases. Genomics technologies have provided rich gene expression datasets for the developing lung that enable systems biology approaches for identifying large-scale molecular signatures within this complex phenomenon. Here, we applied unsupervised principal component analysis on two developing lung datasets and identified common dominant transcriptomic signatures. Of particular interest, we identify an overlying biological program we term "time-to-birth," which describes the distance in age from the day of birth. We identify groups of genes contributing to the time-to-birth molecular signature. Statistically overrepresented are genes involved in oxygen and gas transport activity, as expected for a transition to air breathing, as well as host defense function. In addition, we identify genes with expression patterns associated with the initiation of alveolar formation. Finally, we present validation of gene expression patterns across the two datasets, and independent validation of select genes by qPCR and immunohistochemistry. These data contribute to our understanding of genetic components contributing to large-scale biological processes and may be useful, particularly in animal models of abnormal lung development, to predict the state of organ development or preparation for birth.

Keywords: lung development; microarray; principal component analysis

Development of the mammalian lung follows a complex morphogenetic program whose processes are at least partially conserved across evolution. Research effort over the past few decades has elucidated the essential role of many molecules in the establishment of the mammalian lung (1–4). The initiation of lung formation as a bud off the ventral foregut endoderm ("embryonic" stage of organ development) occurs during the fifth week of gestation in the human, or Embryonic Day (E)9.5 in the mouse, and is promoted by Gli2/3, Shh, and FGF10. Subsequent formation of airway structure of the lung occurs through a process of branching morphogenesis. The regulation of this process, which involves iterative bud elongation and branching, has been extensively studied and is under the control of BMP4, FGF10, Shh, Wnt7b, and additional molecules. Formation of the respiratory portion of the lung, through

CLINICAL RELEVANCE

Lung development is comprised of a series of interrelated morphogenetic events. We identified a molecular signature for birth, consisting of genes involved in this process, based on expression profiling during lung development. These data may be useful in predicting organ development or preparedness for birth.

a process termed alveolarization or alveogenesis, is less well understood. Alveogenesis involves budding and elongation of secondary crests from the primary saccules (ends of conducting airway tubes) as well as reorganization of the vascular capillary bed. We currently appreciate that PDGF-A, FGFR3 and 4, and VEGF are essential for these processes.

A full understanding of lung development (and its pathophysiological perturbations) requires integration of this complex system of morphogenesis and cellular differentiation. For instance, the precise regulation of branch elongation and extension has been shown to require integration of BMP and FGF pathways (5–10). However, nearly all studies to date have involved singular or linear dissection of a single molecule or pathway. Recent work by Whitsett and Matsuzaki used a combination of mouse genetics and functional genomics to integrate individual molecules involved in cellular differentiation and surfactant production into a "circuit" necessary to prepare the lung for the "transition to air breathing" (11).

The recent application of genomics technologies has provided a wealth of data describing biological paradigms. These data are typically distilled to focus upon genes distinguishing one biological state from another. Alternately, time-series data are sometimes used to define individual genes and groups of genes sharing expression patterns in the hope of establishing putative regulatory networks. In either case, the majority of the data is not considered further. Singular value decomposition, or principal components analysis (PCA), is a data-reduction technique that has been widely applied to genome-wide expression profiling datasets. PCA summarizes multifactorial data and reorders the information into major "components" representative of global changes in expression. This technique has been previously been shown to be capable of discovering higher-order biological processes and defining molecular information contributing to them (12, 13).

Here, we have applied PCA to dissect molecular processes contributing to mammalian lung development, using two independent murine gene expression microarray data sets (14, 15). Characterization of each data set, separately, reveals great similarity at the global gene expression level. As anticipated, the greatest influence upon gene expression across mouse lung development is the age of the animal. Strikingly, the second greatest component influencing gene expression in both data sets describes the relative "time-to-birth" for the tissue; that is, the distance in age, before or since, the time of birth. We

(Received in original form January 24, 2008 and in final form May 28, 2008)

* These authors contributed equally to this work.

This work was supported by National Heart, Lung, and Blood Institute contract grant numbers HL71885 (T.J.M., A.T.K.), NS040828 (A.T.K.), and NS047527 (A.T.K.).

Correspondence and requests for reprints should be addressed to Thomas J. Mariani, Ph.D., Division of Neonatology and Center for Pediatric Biomedical Research, University of Rochester, 601 Elmwood Ave., Box 850, Rochester, NY 14642. E-mail: Tom_Mariani@urmc.rochester.edu

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

Am J Respir Cell Mol Biol Vol 40, pp 47–57, 2009

Originally Published in Press as DOI: 10.1165/rcmb.2008-00480C on July 29, 2008

Internet address: www.atsjournals.org

identify groups of genes contributing to this process and those that describe the genomic expression pattern. Interestingly, genes contributing to this process are enriched in functions expected for the transition to air breathing.

MATERIALS AND METHODS

Microarray Data

This study used two previously published, independent microarray datasets of mouse lung development generated from whole lung tissue RNA isolated at various embryonic and postnatal ages. The first dataset, denoted L1, is derived from the A/J strain (14). Gene expression for L1 was measured at six distinct time points—in quadruplicate at E14, E16, E18, and postnatal days (P)7, P14, and P28—using the Affymetrix U74Av2 microarray platform (Affymetrix, Santa Clara, CA), which has 12,488 probes corresponding to 8,809 unique NCBI Entrez genes. The second dataset, denoted L2, is derived from the Swiss-Webster strain (15). Gene expression for L2 was measured at 11 distinct time points—without replicates at E12, E14, E16, E18, P1, P4, P7, P10, P14, and P21 and adult ($\sim$ P30)—using the Affymetrix Mu11K microarray platform, which has 13,179 probes corresponding to 7,202 unique NCBI Entrez genes. There are 6,667 unique Entrez genes in common between L1 and L2; this is the universal set of genes considered in this study.

For each Affymetrix platform, there may be more than one probe reporting the expression level or signal of a given Entrez gene. In such cases, a unique representative probe was selected to represent the Entrez gene using the following repeatability and reproducibility criteria: In L1, we selected the probe with the smallest sum total of intra-time point coefficient of variations across all 6 time points. In L2, there being no replicate sample measurements to use the strategy in L1, we selected the probe with the highest correlation between duplicate time series in a cerebellar development dataset that used the same microarray platform (12).

Normalization and Sample Quality Control

Datasets L1 and L2 are individually normalized by linear regression to the reference sample, which has the highest average linear correlation against all other samples. After normalization, all samples within each dataset have the same average reported expression signal across all probes.

All samples for each dataset were screened for data quality as described in the original publications. Using principal component analysis (described below) of the samples in gene expression space, three samples (E17_1, P14_2, P28_4) in dataset L1 did not co-localize with their co-temporal sample partners and were removed from all subsequent analyses.

Principal Component Analysis

A general microarray gene expression dataset $A = [a_{nm}]$ is a matrix of N genes $\times$ M samples, where typically $N > M$, and a_{nm} is the normalized expression signal of the n -th gene in the m -th sample. Depending on the nature of the investigation, A can be viewed from two distinct perspectives:

- (1) M samples in N -dimensional gene expression space, if one is interested in the relationship between samples (their expression profile) across N genes, or
- (2) N genes in M -dimensional sample space, if one is interested in the relationship between genes (their expression profiles) across M sample conditions.

Principal component analysis: “samples in gene (expression) space.” Algebraically, M objects require at most M independent features for a well-defined characterization. In A , M samples are redundantly determined by N original gene features. The objective here then is to derive a set of $K \leq \min(N, M)$ new features from the N original gene features that characterize the samples nonredundantly. Let S be the $N \times N$ covariance (or more generally, similarity) matrix of A , where the entries s_{ij} denote the covariance (or more generally, measure of similarity) between the i -th and j -th genes across M sample

conditions. The guiding idea is that some of the N original gene features are redundant (correlated or algebraically dependent) across the M samples conditions if and only if S is not a full rank matrix (16), and the eigenvectors of S form a nonredundant (algebraically independent) set of new features for sample characterization. By basic theorems of linear algebra, eigenvectors of S are orthogonal directions of maximal sample variance in A 's original gene feature space.

The m -th sample in the original gene space is $A_m = a_{1m} g_1 + a_{2m} g_2 + \dots + a_{Nm} g_N$ —a vector of length N , where a_{nm} is expression level of gene g_n in A_m , and g_n is a canonical basis element/direction in N dimensions (i.e., g_n is a vector of length N that is 1 at the n -th component and 0 otherwise). Note that the ordered columns vectors g_n form the $N \times N$ identity matrix $I_{N \times N}$. Recall that S is the $N \times N$ covariance matrix of pairwise original gene features g_n s. Eigenvectors of S (the principal components of A) can be computed using singular value decomposition (SVD). Let p_k (or PC_k) denote the k -th principal component (or eigenvector) corresponding to the k -th largest eigenvalue of S , $p_k = c_{1k} g_1 + c_{2k} g_2 + \dots + c_{Nk} g_N$ where c_{nk} is called the *loading* of g_n in p_k and $|c_{nk}|$ signifies the contribution of g_n in p_k . The first principal component p_1 is the direction of maximal sample variance in A . p_2 is the direction of maximal sample variance in the gene feature space orthogonal to space spanned by p_1 . By mathematical induction, p_k is the direction of maximal sample variance in the gene feature space orthogonal to the space spanned by $\{p_1, \dots, p_{k-1}\}$. Let P be the $N \times K$ data matrix with column vectors p_k . Then, the samples represented in the nonredundant new feature space are $A' = P^T A$, where the original data matrix A here has been (row) centered in its feature (here gene) space. A' is a K new feature $\times$ M sample data matrix. P^T denotes the matrix transpose of P . For clarity and contrast with the upcoming principal component analysis (PCA) of “genes in sample space,” we abbreviate the principal components here “gene PC.”

More generally, for any data matrix B whose rows (feature space) are homologous to the rows (feature space) of A , the transformation $B' = P^T B$ is a change of coordinates for columns elements of B from the standard basis into the basis formed by the principal components of A (columns of P). We call this transformation the “projection of B into the principal component space defined by A .”

PCA: “genes in sample space.” PCA of genes in sample space is similar to the foregoing PCA of samples in gene space, except that now the algebraic operations are performed on the transpose of the original data matrix A . The feature space is now defined by samples rather than genes. The n -th gene in the original sample space is $g_n = a_{n1} A_1 + a_{n2} A_2 + \dots + a_{nM} A_M$, where A_m is a canonical basis element/direction in M dimensions that is 1 at the m -th component and 0 otherwise, and the covariance matrix S is $M \times M$ of pairwise original sample features A_m . For clarity and contrast with the earlier PCA of “samples in gene space,” we abbreviate the principal components here “sample PC.”

K-Means Clustering

After PCA of genes in sample space, k -means ($k = 3$) clustering (17) was used to organize the genes in their sample principal components 1 and 2 representation (sample PC1 and PC2) into three disjoint sets of “largely correlated” lung development expression profiles. The choice of $k = 3$ was based on the visual observation of three regions of increased gene density in their sample PC1 versus PC2 plot. To assess the robustness of resulting gene clusters—specifically, to investigate whether any two genes that belong to a common cluster for initial centroids (c_0, c_1, c_2) remain in a common cluster for another different set of initial centroids (c'_0, c'_1, c'_2)— k -means clustering was performed over 10,000 times, each with a random set of three different initial centroids.

Correlations

Standard linear (Pearson) and rank (Spearman) correlations were used as two distinct measures of similarity between two variables (e.g., gene expression or sample profile vectors) (18). Linear correlation assesses how well a linear function describes the relationship between the two variables, whereas rank correlation is a nonparametric measure of correlation assessing how well a monotonic function describes this relationship (i.e., two rankings of matching sets of objects).

Identification of Stage-Specific Genes

We identified probes/genes that were uniquely and significantly over-expressed in each mouse lung development stage relative to the other stages. The pseudoglandular stage was represented by E14 samples ($n = 4$), canalicular by E17 ($n = 3$), saccular by E19 ($n = 4$), and alveolar by P7 ($n = 4$) and P14 ($n = 3$). The expression difference for each probe between stages was computed using a geometric average fold method (12).

Time/Stage of Lung Development and Time-to-Birth Profiles for Computing Correlations

We assume birth to be at E19.5. In dataset L1, the time points can be represented as the vector (E14, E17, E19, P7, P14, P28), which is correlated with (0, 3, 5, 12, 19, 33). The time-to-birth profile is $\text{Abs}((0, 3, 5, 12, 19, 33) - 5.5) = (5.5, 2.5, 0.5, 6.5, 13.5, 27.5)$. In dataset L2, the vector of time points is (E12, E14, E16, E18, P1, P4, P7, P10, P14, P21, P30), which is correlated with (0, 2, 4, 6, 8, 11, 14, 17, 21, 28, 37). The time-to-birth profile is $\text{Abs}((0, 2, 4, 6, 8, 11, 14, 17, 21, 28, 37) - 7.5) = (7.5, 5.5, 3.5, 1.5, 0.5, 3.5, 6.5, 9.5, 13.5, 20.5, 29.5)$.

Ontology Assessment

Functional classification of genes was performed using DAVID 2.007/EASE (<http://david.abcc.ncicrf.gov>, database version March 24, 2008) (19). Entrez GeneIDs for the selected genes were used as the input list, while EntrezIDs for all 6,667 genes included in both datasets served as the background set. Gene Ontology (GO) categories with an EASE score (modified Fisher exact P value) of less than 0.05 were termed significantly overrepresented.

Quantitative Real-Time PCR

Quantitative real-time PCR (qPCR) was performed with custom-designed primers using SYBR Green chemistry on a Stratagene MX3000P (Stratagene, Cedar Creek, TX) essentially as previously described (20). Sequences for the primers used are: Tcf21, 5'-GCTCCCTCAGC GATGTAGAA-3', 5'-GAGTCACAGTCCAGCATCTCC-3'; Meox1, 5'-CTGAGCGGCAGGTCAAAGTC-3', 5'-GTTTCCACTTCATCC TCCGG-3'; Igfbp2, 5'-CAAGTCAGGCATGAAGGAGCT-3', 5'-GCC GGTGCTGTTCATTGA-3'; Lama4, 5'-GACACGTGACCGACATG AAC-3', 5'-CCTTGTTCTGAGGAGGTGGA-3'; Gata6, 5'-AGATGA ATGGCCTCAGCAGG-3', 5'-CAAGCCGCCGTGATGAA-3'.

Immunofluorescence Staining

Immunostaining was performed on formalin-fixed, paraffin-embedded lung tissue sections from wild-type mice essentially as previously described (21, 22). At least three animals were studied at each time point. Detection of IGFBP2 used goat polyclonal anti-mouse Igfbp2 clone M18 (Santa Cruz Biotech, Santa Clara, CA) at a 1:200 dilution and was visualized with Alexa Fluor 488 secondary antibody. Detection of Cnn3 used rabbit polyclonal anti-peptide serum (Abcam Inc., Cambridge, MA) at a 1:200 dilution and was visualized with Alexa Fluor 594 secondary antibody. All slides were counterstained with DAPI to visualize nuclei. Control slides were stained with either secondary antibody alone, purified IgG or preimmune serum.

RESULTS

Identifying Biological Correlates to Dominant Variances within the Developing Lung Transcriptome

We studied two independent gene expression microarray data sets describing mouse lung development, generated on two distinct oligonucleotide-based platforms: Affymetrix Mu11K and Mu74Av2, respectively. For comparison purposes, we restricted the datasets to a set of shared probes representing 6,667 unique genes (see MATERIALS AND METHODS). The first dataset, denoted L1, includes quadruplicate measurements of lung development at six distinct time points: E14, E16, E18, P7, P14, and P28. The second dataset, denoted L2, measured gene expression at 11 distinct time points without replicates: E12, E14, E16, E18, P1, P4, P7, P10, P14, P21, and adult ($\sim$ P30).

PCA was used to study the transcriptome-scale variation (of 6,667 genes) in L1 with linear correlation as a measure of similarity between the samples. This is an unsupervised analysis, as the sample or phenotype labels are not input parameters. Samples are viewed as points in an N -dimensional space defined by $n = 6,667$ gene expression features that can then be summarized into a smaller number of dimensions (1–3) that are correlated with the directions of maximal sample-to-sample variation in the entire dataset. These directions of maximal sample variation are called gene principal components (denoted gene PCs). As shown in Figure 1A, the dominant component of sample variance—that is, the first “gene space” principal component (gene PC1, representing 61.93% of global sample variance)—is correlated with time, or stage of lung development, with a rank correlation $r = 0.9705$ ($P < 0.001$). Interestingly, Scgb1a1 (Utg/CCSP) and Sftpc (SPC) were the greatest contributors to gene PC1, with loadings of 0.3595 and 0.3498, respectively—more than twice the next four highest contributors: Hbb-y (PC1 loading -0.1763), H19 (-0.1620), Ager (0.1610), and Eln (0.1297). Ribosomal proteins also were among the highest contributors to gene PC1 accounting for 38 of the top 100 most negative gene PC1 loading genes. The second greatest component of sample variance (gene PC2, representing 14.87% of global sample variance) is correlated with the distance in sample age to the time of birth, defined as E19.5, with a rank $r = 0.7242$ ($P < 0.001$). The five highest contributors to gene PC2 are: Hbb-b1 (gene PC2 loading 0.3265), Eln (-0.2706), 1200016E24Rik (-0.2074), Hbb-y (0.1909), and Sprr1b (0.1520). Ribosomal proteins comprise 29 of the top 100 most negative gene PC2 loading genes.

To validate that these observed gene PC1–2 biological correlates were not due to technical attributes within L1, we analyzed the second independent mouse lung development microarray dataset L2 using a similar approach. First, we projected the L2 samples into the gene principal component space defined by L1. That is, L2 samples were projected in a 6,667-gene expression space where each gene is weighted by its contribution to sample variance during lung development as described by L1. As shown in Figure 1B, the gene PC1 coordinates of these projected L2 samples are strongly correlated with age (rank $r = 0.9636$; $P < 0.001$), and their gene PC2 coordinates are correlated with time-to-birth (rank $r = 0.6697$; $P = 0.02$). Second, to ensure that we have not artificially imposed the transcriptome-scale dynamics of lung development from L1 onto L2 in the preceding investigation, we performed PCA on L2 independently. As shown in Figure 1D, we find that the gene PC1 coordinates (59.00% of global sample variance) of L2 samples are also correlated with age (rank $r = 0.9636$; $P < 0.001$). Gene PC2 coordinates (gene PC2 representing 17.31% of global sample variance) of L2 are also correlated with time-to-birth (rank $r = 0.5285$; $P = 0.09$). These data reveal that, at the transcriptome scale, the most important component of gene expression variance across lung development is age and the second most important component is correlated with the time-to-birth, independent of dataset, strain, and assaying platform.

A Molecular Signature for Time-to-Birth in the Developing Mouse Lung

Here, PCA was used to study the variation in lung development expression profiles among the 6,667 genes of L1. Genes are viewed as points in an M -dimensional space defined by $M = 21$ sample features—a “genes in sample space” perspective—that can then be summarized into a smaller number of dimensions (1–3) that correspond to the directions of maximal gene-to-gene variation in the entire dataset (denoted sample PCs). In Figure

Figure 1. Principal component analysis (PCA): “samples in 6,667-gene expression space” perspective in the gene PC1 versus gene PC2 plane. (A) PCA of dataset L1 samples. *Magenta*, embryonic; *green*, postnatal. (B) Dataset L2 samples (“X”) projected into dataset L1 gene principal components. *Dark magenta*, embryonic; *dark green*, postnatal. (C) Median gene PC2 (gray) and gene PC1 (cyan) coordinate profiles of samples. (D) PCA of dataset L2 samples (“O”). *Magenta*, embryonic; *green*, postnatal.

2A, each *point* represents a single gene, with its location defined by the two most dominant sample principal components (sample PC1 representing 31.35% of global gene variation, sample PC2 representing 23.72% of global gene variation). The gene’s coordinates on the disc-like “sample space” PC1 versus PC2 plot reflects that gene’s lung development profile. This principal component representation is continuous in the sense that two genes that are close to one another in this sample PC space will have highly linearly correlated profiles. Again, this is an unsupervised analysis as the gene labels are not input parameters.

By visual inspection, we observed three distinct regions of high gene density near the disc’s periphery at approximately 2,

4, and 11 o’clock. We performed a k-means clustering (with $k = 3$) to organize the 6,667 genes into three gene clusters of genes denoted E0 ($n = 2,735/41.02\%$ /gray), E1 ($n = 1,949/29.23\%$ /red), and E2 ($n = 1,983/29.74\%$ /blue). The representative profile for each cluster is shown in Figures 2B to 2D. As a group, E1 (red) represents genes that are more highly expressed early in lung development than later, whereas E2 (blue) are genes that are more highly expressed later in lung development than earlier. Genes in E0 (gray) have more heterogeneous development profiles.

The data presented in Figure 1 reveal gene PC2 to be correlated with the sample age relative to its time-to-birth. As each gene PC is a linear combination of the original 6,667 genes,

Figure 2. Principal component analysis (PCA): “genes in sample space” perspective for L1 dataset in the sample PC1 versus sample PC2 plane. (A) PCA of 6,667 dataset L1 genes partitioned by k-means clustering into three disjoint sets: E0 (gray), E1 (pale red), and E2 (pale blue). Each *dot* marks a gene, and its coordinates in the disk-like scatter plot generally reflect its lung development profile. (B) Standardized time expression profiles of E0 genes. *Black line*, average profile. (C) Standardized time expression profiles of E1 genes. *Red line*, average profile. (D) Standardized time expression profiles of E2 genes. *Blue line*, average profile. For Panels B–D, *x*-axis represents age and *y*-axis represents normalized signal intensity.

we examined the lung development profiles of the 200 genes with the greatest contribution (highest loadings) to gene PC2: the 100 genes with the most positive (≥ 0.012125) and the 100 genes with the most negative (≤ -0.035349) gene PC2 loadings. Figure 3A shows the sample PC profiles for these 200 gene PC2 genes, relative to all 6,667 genes, with genes having positive loadings in *magenta* and those having negative loadings in *green*. Their corresponding expression profiles are shown in Figures 3B and 3C. Note that due to the unsupervised and linear combinatorial nature of each PC, the sample expression profiles of individual genes need not necessarily be correlated with the gene PC2 sample coordinates as shown in Figure 1C. That is, they do not necessarily have minimum or maximum expression at the time of birth. Likewise, the direction of contribution (positive/negative loadings) to gene PC2 is not necessarily related to the expression profile of that gene at birth. However, it is the linear combination of these genes that is correlated with the age of the sample relative to birth.

We performed a gene ontology (GO) assessment of these 200 genes with the highest loadings in gene PC2 (see Table E1 in the online supplement). This list showed a significant overrepresentation (EASE score < 0.05) of genes whose ontologic attributes are relevant to the transition to air breathing, such as oxygen transport and binding (4.14- to 20.51-fold overrepresented), and immunologic and antigen processes (3.58- to 11.29-fold). Additional ontologic categories pertaining to the hemoglobin (2.85-22.58-fold) and ribosomal (4.85- to 19.51-fold) complex, macromolecule metabolism (1.24- to 3.25-fold), and nitric oxide synthesis (9.80- to 23.42-fold) were also overrepresented. There was good agreement between the gene PC2 highest loading genes in both datasets L1 and L2 with respect to the categories overrepresented.

For a more direct, supervised approach motivated by the time-to-birth association of the gene PC2 coordinates of L1 samples, we next identified genes whose sample lung development expression profiles were correlated (positively and negatively) with their gene PC2 coordinates. For each of the 6,667 genes, we computed the linear correlation of its sample development profile against the constant gene PC2 coordinates (Figure 1C). The 100 most positively correlated genes ($r \geq 0.70058$) have a lung development expression profile with a global minimum around E19.5; whereas the 100 most negatively correlated genes ($r \leq -0.76311$) have a global maximum around E19.5. To visualize their sample PC profiles among all 6,667 genes, these highly correlated genes are marked on the sample PC1 versus PC2 plot in Figure 4A. Their lung development expression profiles are shown in Figures 4B and 4C. There were only 19 genes in common between this list of correlated genes and the earlier list of 200 genes with the highest loadings in gene PC2, indicating substantial distinctions in inference resulting from unsupervised (PCA) versus supervised (correlated with given pattern) analysis.

To further address the value and independence of our unsupervised approach, we completed a supervised analysis of the data to identify stage-specific patterns of expression (Table E2). In particular, we examined the intersection between the “stage-specific” genes uniquely overexpressed at E19 ($n = 131$) and the genes ($n = 200$) with the greatest PC2 loadings (gene PC2 being correlated with time-to-birth). Only 26 genes were in common in this analysis, demonstrating the distinct nature of the results from supervised and unsupervised analyses.

An Association between Gene PC3 and Alveogenesis

We have observed that the dominant variance structures or PC's in datasets L1 and L2 appear to be correlated with developmental age and the time-to-birth. As seen in Figure

5A, gene PC3 (capturing 6.56% of global sample variance) of dataset L1 is correlated with the initiation of alveolar formation at P7: P7 samples lie at the extreme positive end of the gene PC3 axis. As was done for gene PC2, we examined the lung development profiles of the 200 genes with the greatest contribution (highest loadings) to gene PC3 (gene PC3 loading ≥ 0.033045 or ≤ -0.016221). A sizable proportion had developmental time profiles with a global maximum at P7, and a local maximum at E14 (Figure 5D), essentially the gene PC3 coordinate profile of L1 samples (Figure 7B). Thirty-three of these 200 top gene PC3 contributors, 33 (16.5%)—a significant proportion, $P < 10^{-12}$ two-tail Fisher exact test—were also among the top 100 most positively ($r \geq 0.61400$) and 100 most negatively ($r \leq -0.57638$) correlated to the gene PC3 coordinate profile. Six of these genes are highly correlated against the gene PC3 coordinate profile (high $r_{L1,PC3}$) and have consistent developmental profiles across datasets L1 and L2 (Figure E1): Cnn3 ($r_{L1,PC3} = 0.81567$), Col5a2 ($r_{L1,PC3} = 0.74219$), Lgals1 ($r_{L1,PC3} = 0.79584$), Rbp1 ($r_{L1,PC3} = 0.64497$), Oaz1 ($r_{L1,PC3} = 0.62631$), and Igfbp6 ($r_{L1,PC3} = -0.64323$).

Profile Consistency and Experimental Validations of the Developing Mouse Lung Transcriptome

Gene-by-gene, we investigated the consistency of expression profiles of the 6,667 genes across datasets L1 and L2. Recall that L1 is composed of replicated transcriptome profiles at E14, E17, E19, P7, P14, and P28, whereas L2 is nonreplicated transcriptome profiles at E12, E14, E16, E18, P1, P4, P7, P10, P14, P21, and adult ($\sim P30$). To render L2 comparable to the time points of L1, we considered the following five time point sample profiles for each gene in L2: E14, median (E16, E18), P7, P14, median (P21, adult), relative to the E14, E17, P7, P14, P28 median time points in L1. A histogram of the rank correlation for genes across datasets L1-L2 (rank r_{L1-L2}) is shown in Figure 6. The median rank correlation was 0.3000, indicating that the majority of genes had consistent (positively rank-correlated) lung development expression profiles.

Quantitative PCR (qPCR) was used for independent developmental expression profile validation at five time-points: E16, E18, P1, P7, and P14. Here, we spot-picked five genes whose sample profiles were among the top 200 most correlated against the constant gene PC2 sample coordinates in L1 (high $r_{L1,PC2}$), but whose cross L1-L2 dataset rank-correlations, as explained in the preceding paragraph, were variable: Tcf21 ($r_{L1,PC2} = -0.9137$, rank $r_{L1-L2} = 1.00$), Igfbp2 ($r_{L1,PC2} = 0.7471$, rank $r_{L1-L2} = 0.70$), Lama4 ($r_{L1,PC2} = -0.8370$, rank $r_{L1-L2} = 0.50$), Gata6 ($r_{L1,PC2} = -0.7919$, rank $r_{L1-L2} = 0.20$), and Meox1 ($r_{L1,PC2} = 0.7310$, rank $r_{L1-L2} = 0.10$). Figure 6 shows the expression profiles of these genes as reported by dataset L1 (*light green*), L2 (*pink*), and qPCR (*dark green*). We observed that the level of concordance between L1 and qPCR reported lung development profiles for these five genes is proportional to the magnitude of the cross L1-L2 dataset rank correlation (rank r_{L1-L2}) value.

Finally, we performed immunofluorescence staining to determine the cell populations responsible for developmental changes in gene expression for two of these genes. Detection of Igfbp2 (high gene PC2 loading gene) was confirmed in the distal alveolar region of the lung (Figure 7). Low levels of expression were observed in the newborn, with peak expression 1 to 2 weeks after birth and returning to lower levels at later time points. This pattern is consistent with the microarray results. Staining revealed a distributed pattern throughout the alveolar wall, lacking a specific cell type-specific pattern. These data indicate that regulation of whole lung Igfbp2 expression

Figure 3. Genes with the highest gene PC2 loadings ("samples in gene space perspective") in dataset L1. (A) Top 100 most positive (magenta "X") and negative (green "X") gene PC2 loading genes highlighted among all 6,667 gene expression profiles. Dark magenta/green mark genes whose expression profiles are additionally among the top 200 most correlated against gene PC2 sample coordinates ("samples in gene space perspective"). (B) Standardized time expression profiles of top 100 most positive loading genes in gene PC2. (C) Standardized time expression profiles of top 100 most negative loading genes in gene PC2. For Panels B and C, x-axis represents age and y-axis represents normalized signal intensity.

levels results from changes in the amount produced by individual cells rather than a change in the population of cells expressing the gene. Detection of acidic calponin (Cnn3; high gene PC3 loading gene) revealed a distinct pattern of expression (Figure 8). Cnn3 was clearly detected in vessels (and, to a lesser extent in the airways) in newborn mice. As expected, expressing cells were restricted to those with a smooth muscle phenotype, being subjacent to cells lining the vascular endothelium (and airway epithelium). At 1 week of age, Cnn3 staining became more prominent, associated with an expansion of vascular smooth muscle. At 2 and 4 weeks of age, Cnn3 staining was

diminished, particularly from vascular structures. This pattern is consistent with the microarray results. Appreciable staining in the alveolar region was not noted. These data indicate that regulation of whole lung Cnn3 expression results from changes in the number of cells expressing this protein in the lung.

DISCUSSION

Principal components analysis is a data reduction strategy that has previously been used to identify dominant variance structures within multifactorial data, and can be used to characterize

Figure 4. Genes whose developmental profiles are most highly correlated to the gene PC2 sample coordinates ("samples in gene space perspective") in dataset L1. (A) Top 100 most positive (magenta "X") and negative (green "X") gene PC2 correlated genes. Dark magenta/green mark genes mark genes that are additionally among the top 200 highest gene PC2 loadings ("samples in gene space perspective") (cf. Figure 3). (B) Standardized time expression profiles of top 100 most positively correlated genes to gene PC2 sample coordinates. (C) Standardized time expression profiles of top 100 most negatively correlated genes to gene PC2 sample coordinates. For Panels B and C, x-axis represents age and y-axis represents normalized signal intensity.

Figure 5. Principal component analysis (PCA): “samples in 6,667-gene expression space” perspective focusing on the gene PC3 axis. (A) PCA of dataset L1 samples. *Magenta*, embryonic; *green*, postnatal. (B) Median gene PC3 (*gray*) and gene PC1 (*cyan*) coordinate profiles of samples. (C) Top 100 most positive (*magenta* “X”) and negative (*green* “X”) Top 100 most positive (*magenta* “X”) and negative (*green* “X”) Top 100 most positive (*magenta* “X”) and negative (*green* “X”) gene PC3 loading genes highlighted among all 6,667 gene expression profiles. *Dark magenta/green* mark genes whose expression profiles are additionally among the top 200 most correlated against gene PC3 sample coordinates (“samples in gene space perspective”). (D and E) Standardized time expression profiles of top 100 most positive (*magenta*) and top 100 most negative (*green*) loading genes in gene PC3.

global biological functions in gene expression microarray data (12, 13, 23, 24). Here, we report applying PCA to microarray datasets describing global gene expression patterns during mouse lung development. This was done to characterize global programs of gene expression, identify and define the molecular bases of defined histologic/physiologic stages, and identify novel genes representing putative biological modules relevant to the lung development process. Although describing the same biological process, the two datasets used for these studies are significantly different in terms of technical attributes including the microarray platform, time points studied, and nature of replication strategy (L1 used technical replicates of samples derived from individuals, L2 used pooling of individual samples). Furthermore, these data are derived from distinct strains of mice. Regardless, we find that the global expression patterns between the two datasets are remarkably similar. The basis of this similarity appears to be due to two major, biologically interpretable influences upon the transcriptome profiles. The first, as defined by gene PC1, is the age of the sample. This is wholly anticipated, as the experimental design was meant to capture differences in gene expression relative to age. Indeed, if gene PC1 sample coordinates were not correlated with age, then the major objective of the studies would be suspect. To our

surprise, we find the second major biological influence upon global sample-to-sample gene expression variation (gene PC2), rather than any particular stage of development, is the relative distance of each sample in days from the time of birth. These observations were clearly evident and independently verified by separate analyses of each dataset. Finally, we observed the third greatest component of variance was associated with the initiation of alveolar formation.

Each gene PC is a composite of 6,667 distinct gene expression patterns, such that profiles of individual genes do not necessarily resemble the pattern described by the gene PC sample coordinates. In addition, membership in each gene PC is not exclusive, meaning that each gene can contribute to multiple gene PCs. We noted that the genes that contributed most to gene PC1 (large gene PC1 loadings), which is correlated with developmental age, were lung-specific genes whose expression is initiated during the developmental program. We investigated gene PC2 in an effort to better define gene expression and genetic programs that may be related to the time-to-birth pattern that reflects a transition to air breathing. We found that a number of biological functions were over-represented, including many intuitive for a transition to air breathing, such as gas transport and oxygen transport.

Figure 6. Cross-dataset (L1-L2) lung development profile consistency and validations. (*Top left quadrant*) Histogram of rank correlations of 6,667 genes across datasets L1-L2 on 5 matched time points. *Red vertical lines* mark various quantile points. (*Small side diagrams*) Standardized expression profiles of five genes (Tcf21, Igfbp2, Lama4, Gata6, and Meox1) as reported in L1 (light green), L2 (pink), and independently assessed via qPCR (dark green). For validation panels, x-axis represents age and y-axis represents normalized signal intensity.

Surfactant production is arguably the most important factor for survival after birth. Surfactant protein B, the major genetic determinant of neonatal respiratory distress, was among the top 200 genes (#190 in gene PC2). Surfactant protein C was not among the greatest contributors to gene PC2, presumably due to its induction before birth (#2 in gene PC1, #3715 in gene PC2). Three of the top six genes contributing to gene PC2, including the gene contributing the most to PC2, encoded components of hemoglobin. Obviously, hemoglobin is essential for breathing and gas transfer. Changes in expression for these genes (Hbb-b1, Hbb-y, Hbb-b2) may reflect a static level of expression coupled with an increase in the red blood cell population within the lung. Alternately, this may reflect an unanticipated, and previously unappreciated, expression of these genes at the RNA level within the lung. The gene that contributed the second most to PC2 in L1 was Eln, the gene encoding the major component of elastic fibers. Elastic fibers are responsible for providing tissue elastic recoil essential for expiration and has previously been shown to be essential for survival at birth (25, 26). However, the complex expression pattern for elastin (Eln) expression during lung development (15) is not correlated with the time-to-birth and thus would not clearly indicate its necessity for the transition to air breathing. The essential role of elastin in neonatal survival, combined with

its complex expression pattern seemingly unrelated to birth, exemplifies the strength of our approach.

To further understand the genomic program for the transition to air breathing, or time-to-birth, we sought to identify individual genes whose expression profiles were indicative of this pattern. We identified a relatively small number of genes whose expression was consistently correlated with gene PC2 sample coordinates across the two datasets, having a pattern of expression that was either maximal or minimal at or near the time of birth. qPCR analysis verified the expression patterns for a number of these genes. Among these 18 genes, some are particularly interesting: Tcf21 (Capsulin/Epicadin/Pod-1) has previously been shown to be essential for lung alveolar development and survival at birth (27). Gata6 is known to be important for lung development, branching morphogenesis and respiratory epithelial cell differentiation (28–31). Npr3 (natriuretic peptide receptor-3) has previously been shown to be rapidly down-regulated after birth and is proposed to maintain the fetal lung in a fluid-filled state (32). Pld1 is a phosphatidylcholine-specific phospholipase expressed in the lung (33, 34) that may be involved in surfactant metabolism.

Another of these 18 genes relates to previous studies characterizing the transcriptional regulation of lung maturation and perinatal adaptation to air breathing; Ppp3cb, CnB1 or

Figure 7. Immunostaining for IGFBP2. Immunofluorescent detection of IGFBP2 during postnatal development was assessed using a goat polyclonal antibody as described in MATERIALS AND METHODS. Shown are representative panels for staining in normal mouse lungs at 1 day, 1 week, 2 weeks, or 4 weeks of age. The *lower right panel* shows a merged image with DAPI counterstaining of nuclei, to allow better visualization of cell-specific staining. Original magnification (all panels) = $\times 400$.

calcineurinB. Whitsett and colleagues have characterized a regulatory network essential for the production of pulmonary surfactant, a critical component of the transition to air breathing (35). This regulatory network involves cooperation between the transcription factors *Titf1*, *Foxa1*, *Foxa2*, and *Cebpa*. Recent work (36) has discovered that perinatal maturation of the lung is principally controlled by calcineurin signaling. Ultimately, under the control of *Ppp3cb*, *Foxa2* promotes surfactant protein, *Abca3*, and *Cebpa* gene expression (37, 38). Our data identifying *Ppp3cb* as a gene contributing to the time-to-birth molecular signature validates the existence of gene expression modules focused upon preparing the mammalian lung for transition to air breathing and identifies potential novel members of this biological program.

In terms of the discovery of PC3 genes, and their relationship to the initiation of alveolar formation, we focused upon a subset of genes whose expression patterns contributed to PC3 loadings, correlated with the PC3 profile and were consistent between the two datasets. This conservative approach was motivated by the limited informativity of PC3 ($\sim 6\%$ variance) and poor resolution of PC3 in L2. Regardless, we identified a set of six genes that met these criteria and whose expression peaks at the initiation of

alveolar formation. Among these, *Lgals1* (galectin-1) has recently been shown to have the expression pattern described by the array data, and to be expressed within secondary alveolar septae uniquely being formed at this stage (39). In addition, spatial and temporal changes in the expression of *rbp1* may participate in retinoid-dependent alveolar formation (40–46).

While our study focuses upon the consistencies between the two datasets analyzed, we have also performed an unsupervised assessment of correlation in sample expression patterns for the 6,667 genes common to them (Figure 6). These data identify a wide range of reliability among individual profiles across the two datasets, with some genes displaying very similar patterns and others displaying almost opposing patterns ($r < 0$). While we have not directly investigated the nature of the discrepancies, possible sources of variability include measurement precision (47), probe sequence errors (48, 49), and experimental technical differences. Whether any of these inconsistencies are due to *bone fide* differences in individual gene expression patterns between distinct strains of mice will require further evaluation. Additional limitations of the current study relate to “sampling.” First, the applicability of these findings is limited by the availability of lung tissue. Whether a similar influence

Figure 8. Immunostaining for CNN3. Immunofluorescent detection of acidic calponin during postnatal development was assessed using a rabbit polyclonal antibody as described in MATERIALS AND METHODS. Shown are representative panels for staining in normal mouse lungs at 1 day, 1 week, 2 weeks, or 4 weeks of age. The *lower right panel* shows a merged image with DAPI counterstaining of nuclei, to allow better visualization of cell-specific staining. Original magnification (all panels) = $\times 400$.

of time-to-birth upon gene expression would be captured in samples obtained by less invasive means, such as peripheral blood or intra-airway cells, is unclear. Second, the resolution of the datasets, and our subsequent results, are limited by the low frequency of time points studied. Efforts to define regulatory networks or modules are severely hampered by the complexities in gene expression changes that are likely to occur in the days between the intervals measured in the current datasets.

In summary, we have found that preparation of the lung for the transition to air breathing has a dominant effect on global gene expression patterns during mouse lung development. We have identified genes contributing to this regulatory program and shown that as a group they are involved in obvious (oxygen/gas transport, morphogenesis, host defense/immunity) and unexpected (nitric oxide synthesis, and ribosomal and hemoglobin complex) biological functions. These genes are informative in terms of defining the relative state of mouse lung tissue in terms of its preparedness for the transition to air breathing, and thus represent a molecular signature for time-to-birth. A complete list of the PCA data described in this report, along with cross-platform gene-specific correlations, is provided in Table E3.

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.

References

- Warburton D, Lee M. Current concepts on lung development. *Curr Opin Pediatr* 1999;11:188–192.
- Cardoso W. Lung morphogenesis revisited: old facts, current ideas. *Dev Dyn* 2000;219:121–130.
- Cardoso WV, Lu J. Regulation of early lung morphogenesis: questions, facts and controversies. *Development* 2006;133:1611–1624.
- 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.
- Sutherland D, Samakovlis C, Krasnow M. Breathless encodes a Drosophila FGF homolog that control tracheal cell migration and the pattern of branching. *Cell* 1996;87:1091–1101.
- Bellusci S, Grindley J, Emoto H, Itoh N, Hogan B. Fibroblast growth factor 10 (FGF10) and branching morphogenesis in the embryonic mouse lung. *Development* 1997;124:4867–4878.
- Park W, Miranda B, Lebeche D, Hashimoto G, Cardoso W. FGF-10 is a chemotactic factor for distal epithelial buds during lung development. *Dev Biol* 1998;201:125–134.
- Bellusci S, Henderson R, Winner G, Oikawa T, Hogan B. Evidence from normal expression and targeted misexpression that bone morphogenetic protein (BMP-4) plays a role in mouse embryonic lung morphogenesis. *Development* 1996;122:1693–1702.
- Weaver M, Dunn N, Hogan B. Bmp-4 and FGF-10 play opposing roles during lung bud morphogenesis. *Development* 2000;127:2695–2704.
- Bragg AD, Moses HL, Serra R. Signaling to the epithelium is not sufficient to mediate all of the effects of transforming growth factor beta and bone morphogenetic protein 4 on murine embryonic lung development. *Mech Dev* 2001;109:13–26.
- Whitsett JA, Matsuzaki Y. Transcriptional regulation of perinatal lung maturation. *Pediatr Clin North Am* 2006;53:873–887. (viii).
- 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.
- Misra J, Schmitt W, Hwang D, Hsiao LL, Gullans S, Stephanopoulos G, Stephanopoulos G. Interactive exploration of microarray gene expression patterns in a reduced dimensional space. *Genome Res* 2002;12:1112–1120.
- 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.
- 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.
- Johnson RA, Wichern DW. Applied multivariate statistical analysis, 5th ed. Upper Saddle River, NJ: Prentice Hall; 2002.
- Duda RO, Hart PE, Stork DG. Pattern classification, 2nd ed. New York: Wiley; 2001.
- Arbeitman MN, Furlong EE, Imam F, Johnson E, Null BH, Baker BS, Krasnow MA, Scott MP, Davis RW, White KP. Gene expression during the life cycle of *Drosophila melanogaster*. *Science* 2002;297:2270–2275.
- Hosack D, Jr GD, Sherman B, Lane H, Lempicki R. Identifying biological themes within lists of genes with EASE. *Genome Biol* 2003;4:4.
- Arikan M, Shapiro S, Mariani T. Induction of macrophage metalloelastase (MMP12) expression by statins. *J Cell Physiol* 2005;204:139–145.
- Mariani TJ, Crouch E, Roby JD, Starcher B, Pierce RA. Increased elastin production in experimental granulomatous lung disease. *Am J Pathol* 1995;147:988–1000.
- Simon DM, Arikan MC, Srisuma S, Bhattacharya S, Tsai LW, Ingenito EP, Gonzalez F, Shapiro SD, Mariani TJ. Epithelial cell PPAR[gamma] contributes to normal lung maturation. *FASEB J* 2006;20:1507–1509.
- Holter N, Mitra M, Maritan A, Cieplak M, Banavar J, Fedoroff N. Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc Natl Acad Sci USA* 2000;97:8409–8414.
- Wen X, Fuhrman S, Michaels G, Carr D, Smith S, Barker J, Somogyi R. Large-scale temporal gene expression mapping of central nervous system development. *Proc Natl Acad Sci USA* 1998;95:334–339.
- Mariani T, Pierce R. Development of lung elastic matrix. In: Gaultier G, Bourbon J, Post M, editors. Lung development. New York: Oxford University Press; 1999. pp. 28–45.
- Mariani TJ, Sandefur S, Pierce RA. Elastin in lung development. *Exp Lung Res* 1997;23:131–145.
- Quaggin SE, Schwartz L, Cui S, Igarashi P, Deimling J, Post M, Rossant J. The basic-helix-loop-helix protein pod1 is critically important for kidney and lung organogenesis. *Development* 1999;126:5771–5783.
- Keijzer R, van Tuyl M, Meijers C, Post M, Tibboel D, Grosveld F, Koutsourakis M. The transcription factor GATA6 is essential for branching morphogenesis and epithelial cell differentiation during fetal pulmonary development. *Development* 2001;128:503–511.
- Liu C, Ikegami M, Stahlman MT, Dey CR, Whitsett JA. Inhibition of alveolarization and altered pulmonary mechanics in mice expressing GATA-6. *Am J Physiol Lung Cell Mol Physiol* 2003;285:L1246–L1254.
- Liu C, Morrissey EE, Whitsett JA. GATA-6 is required for maturation of the lung in late gestation. *Am J Physiol Lung Cell Mol Physiol* 2002;283:L468–L475.
- Yang H, Lu MM, Zhang L, Whitsett JA, Morrissey EE. GATA6 regulates differentiation of distal lung epithelium. *Development* 2002;129:2233–2246.
- D'Angelis CA, Nickerson PA, Ryan RM, Swartz DD, Holm BA. C-type natriuretic peptide and its receptor are downregulated in pulmonary epithelium following birth. *Histochem Cell Biol* 2006;126:317–324.
- Wang L, Cummings R, Usatyuk P, Morris A, Irani K, Natarajan V. Involvement of phospholipases D1 and D2 in sphingosine 1-phosphate-induced ERK (extracellular-signal-regulated kinase) activation and interleukin-8 secretion in human bronchial epithelial cells. *Biochem J* 2002;367:751–760.
- Wang L, Cummings R, Zhao Y, Kazlauskas A, Sham JK, Morris A, Georas S, Brindley DN, Natarajan V. Involvement of phospholipase D2 in lysophosphatidate-induced transactivation of platelet-derived growth factor receptor-beta in human bronchial epithelial cells. *J Biol Chem* 2003;278:39331–39940.
- Maeda Y, Dave V, Whitsett JA. Transcriptional control of lung morphogenesis. *Physiol Rev* 2007;87:219–244.
- 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.
- Martis PC, Whitsett JA, Xu Y, Perl AK, Wan H, Ikegami M. C/EBPalpha is required for lung maturation at birth. *Development* 2006;133:1155–1164.
- Wan H, Xu Y, Ikegami M, Stahlman MT, Kaestner KH, Ang SL, Whitsett JA. Foxa2 is required for transition to air breathing at birth. *Proc Natl Acad Sci USA* 2004;101:14449–14454.
- Foster JJ, Goss KL, George CL, Bangsund PJ, Snyder JM. Galectin-1 in secondary alveolar septae of neonatal mouse lung. *Am J Physiol Lung Cell Mol Physiol* 2006;291:L1142–L1149.
- Hind M, Corcoran J, Maden M. Temporal/spatial expression of retinoid binding proteins and RAR isoforms in the postnatal lung. *Am J Physiol Lung Cell Mol Physiol* 2002;282:L468–L476.
- Massaro G, Massaro D. Retinoic acid treatment partially rescues failed septation in rats and in mice. *Am J Physiol Lung Cell Mol Physiol* 2000;278:L955–L960.

42. Massaro GD, Massaro D. Postnatal treatment with retinoic acid increases the number of pulmonary alveoli in rats. *Am J Physiol* 1996;270:L305–L310.
43. Massaro GD, Massaro D. Retinoic acid treatment abrogates elastase-induced pulmonary emphysema in rats. *Nat Med* 1997;3:675–677.
44. Massaro GD, Massaro D, Chambon P. Retinoic acid receptor- α regulates pulmonary alveolus formation in mice after, but not during, perinatal period. *Am J Physiol Lung Cell Mol Physiol* 2003;284:L431–L433.
45. McGowan S, Jackson SK, Jenkins-Moore M, Dai HH, Chambon P, Snyder JM. Mice bearing deletions of retinoic acid receptors demonstrate reduced lung elastin and alveolar numbers. *Am J Respir Cell Mol Biol* 2000;23:162–167.
46. McGowan SE. Contributions of retinoids to the generation and repair of the pulmonary alveolus. *Chest* 2002;121(5, Suppl):206S–208S.
47. Mariani TJ, Budhraj V, Mecham BH, Gu CC, Watson MA, Sadovsky Y. A variable fold change threshold determines significance for expression microarrays. *FASEB J* 2003;17:321–323.
48. Mecham BH, Klus GT, Strovel J, Augustus M, Byrne D, Bozso P, Wetmore DZ, Mariani TJ, Kohane IS, Szallasi Z. Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Res* 2004;32:e74.
49. Mecham BH, Wetmore DZ, Szallasi Z, Sadovsky Y, Kohane I, Mariani TJ. Increased measurement accuracy for sequence-verified microarray probes. *Physiol Genomics* 2004;18:308–315.