Supplement to Integrating multi-omics longitudinal data to reconstruct networks underlying lung development

Contents

Supporting methods

Preparation of Lung tissue for Laser Capture Microdissection (LCM)

All animal procedures were approved by the Institutional Animal Care and Use Committee (IACUC) at the University of Alabama at Birmingham. The lung tissue used in this study was obtained from pups born to C57BL/6 timed-pregnant females purchased from Jackson Laboratories (Bar Harbor, ME). Mice lung samples were collected over the first 28 days of lung development (every 12h for the first 14d, then every 24h).

Human lung samples were supplied by the LungMAP Human Tissue Core (Biorepository for Investigation of Neonatal Diseases of Lung-Normal [BRINDL-NL]; University of Rochester Medical Center). The Human Tissue Core procured, processed, deposited, and distributed normal late fetal, neonatal, and early childhood human lung tissue and dissociated cells for the LungMAP project working with United Network for Organ Sharing and their network partners. Transplantation quality (but non-transplantable, with a warm ischemia time <12 hours) tissues were collected through the International Institute for the Advancement of Medicine and the National Disease Research Interchange, organizations that link donors to the scientific community. As post-mortem lungs were used, this project was classified as "Not Human Subjects Research" by the NIH. Eligible individuals were liveborn, age 20 weeks gestation to 10 years old, with known time of death and minimal warm ischemia time preferred < 6hours, max 12hours. These infants/children most commonly died of lethal brain or cardiac disease, or trauma for older children. Exclusions were unknown ischemia time or warm ischemia time >12h, primary lung injury (e.g smoke inhalation, pneumonia, drowning) or prolonged ventilation. All lungs were also evaluated by radiography and histology, and abnormal specimens were not used.

Lung tissue specimens were immediately embedded in Optimal Cutting Temperature (OCT, Tissue-Tek#4583) after dissection, snap frozen in liquid N_2 and transferred at - 80°C for storage. Tissue sections (16 µm) were cut with a cryostat (Thermo scientific) onto membrane-glass slides (Arcturus PEN#LCM0522) at −25 °C. The OCT slides were washed and briefly dehydrated prior to LCM using Arcturus® HistoGene® LCM Frozen Section Staining Kit. Using the ArcturusXT system, lung parenchymal samples were collected in CapSure® LCM Caps (#LCM0211) and stored in lysis buffer (Invitrogen) at - 80°C or flash frozen in liquid N_2, and then shipped in bulk on dry ice for further downstream analysis. More specifically, the IR laser was used to select the alveolar areas, and the UV laser was used to cut widely around the selected areas to avoid large blood vessels and bronchi. Adjacent histological sections of each of the lungs used as biological replicates were used for LCM and then subjected to the analytical methods (for mRNA, miRNA, proteomic, and DNA methylation) to ensure that data from all methods represented similar regions of the lung. The human lung samples were collected using the LCM procedure described above at different time points during lung development: Early (1d to 1 month old), Mid A (2 months to 1 years old), Mid B (1 year to 2 years old), Mid C (3 to 4 years old), and Late (8 to 9 years old).

mRNA analysis

Total RNA was extracted according to the miRneasy Micro Kit (Qiagen217084) according the manufacturer's protocol with some modification to gain the best recovery from LCM samples. After evaluation of RNA Integrity Number (RIN), 100 ng input were used for library preparation by Illumina® TruSeq® Stranded Total RNA Sample Preparation kits with Ribo-Zero according manufacturer's protocols. DNA libraries were sequenced on Illumina Hi-Seq 2500 generating paired-end 75 bp reads with a mean of 20 M reads per sample. Fastq files after quality control for adapter contamination and trimming, were mapped on mouse genome by most updated version of STAR software. FPKM matrix was generated by cufflinks for data analysis.

Human samples: 20 ng RNA input was used for RNA sequencing per sample, using Ion AmpliSeq™ Transcriptome Human Gene Expression kit according the manufacturer's protocol. The cDNA was then amplified and barcoded using the Ion Xpress™ RNA-Seq Barcode 1-16 Kit (Life Technologies 4475485). The Ion PI™ Chip Kit v2 BC (Life Technologies 4484270) was then loaded using the Ion Chef™ System (Life Technologies 4484177) with the Ion PI™ IC 200 Kit (Life Technologies 4488377). Afterwards, the chips were run on the Ion Proton™ System for Next-Generation Sequencing (Life Technologies 4476610) using the Ion PI™ IC 200 Kit (Life Fastq files after quality control for adapter contamination and trimming, were mapped on human genome by combinations of Tophat and Bowtie2 software. FPKM matrix was generated by cufflinks for data analysis.

DNA methylation analysis

Mouse genomic DNA from lung tissues was isolated from Laser Capture Microdissected (LCM) samples at different time points during lung development: saccular stage (P0.5, and P2.5) and alveolar stage (P7, P10, P19 and P28). DNA was isolated using the ZR Genomic DNA-Tissue MicroPrep kit (Zymo Research) according to the specific protocol for extraction of genomic DNA from solid tissues. Incubation with Digestion buffer and proteinase K was done overnight at 55^{o}C in inverted tubes. DNA was eluted with Elution Buffer pre-equilibrated to 60-70^{o}C. DNA concentration was determined using the Qubit dsDNA HS assay kit (Thermo Scientific).

Three biological samples per time point were analyzed by MeDIP-seq. Briefly, DNA (gDNA) was sonicated in a Bioruptor Pico (Diagenode) at 2.2 ng/µL for 13 cycles of 30s on/ 30s off to 100-500 bp, with a mean fragment size of 185 bp. Fragment length distribution was assessed using an Agilent 2100 Bioanalyzer and high-sensitivity DNA chips. Sonicated gDNA (in the range of 88–125 ng) was further used for end-repair and adaptor ligation employing the NEBNext Ultra DNA Library Prep Kit for Illumina (New England Biolabs).

The reaction mix was then purified using AxyPrep MagTM PCR Clean-up (Corning Life Sciences). Adaptor-ligated DNA was spiked with a 5mC DNA standard (Zymo Research, catalog D5405-2) in the ratio 1:100 spike: genomic DNA. After 10 minutes of denaturation at 95^{o}C, spiked adaptor ligated samples were cooled in an ice bath for 10 minutes, and 1/10 of the reaction volume was stored as Input sample at 4^{o}C. MeDIP reaction included both 5-Methylcytosine (Active Motif, catalog: 39649) and bridging antibody (Active Motif, catalog 53017) each at 2.5 ug of antibody/ 1ug adapter ligated DNA. Effectively, between 71.3 and 100 ng of spiked adaptor-ligated DNA was used in the MeDIP process. Antibody incubation was performed at 4°C overnight in the presence of Protein G magnetic beads (Active Motif, catalog: 53033).

The DNA:antibody:beads complex was magnetically collected, and washed three times with 1X IP buffer [10mM sodium phosphate buffer (pH 7), 140mM NaCL, 0.05% Triton X-100] . Beads were then incubated with proteinase K (0.3mg/ml) for 2 hours at 50^{o}C and eluates (antibody-enriched fraction) and Input samples were purified using the ChIP DNA Clean & Concentrator (Zymo Research, catalog: D5205). Recovery and enrichment was evaluated by qPCR using primer sets specific for mouse methylated and unmethylated regions (1), and the spike-in control (Forward: AGGTGGAGGAAGGTGATGTC, Reverse: 5' ATAAACCGAACCGCTACACC).

To complete library preparation, antibody-enriched DNA and Input samples were PCR-amplified using standard Illumina index primers (New England Biolabs). PCR products were then purified using AxyPrep MagTM PCR, and eluted with 10mM TRIS-HCl pH 8.0. Library concentration was measured by Qubit dsDNA HS assay kit (Thermo Scientific), and size selected using the PippinHT platform in 2 % agarose gel (Sage Sciences). Size distribution was analyzed in an Agilent 2100 Bioanalyzer with high-sensitivity DNA chips. Libraries were then uploaded onto an Illumina HiSeq4000 platform at the IGM Genomics Center (University of California San Diego) for ultra-deep sequencing with single reads of 50 bases. Forty million reads per sample were sequenced. Novogene Corporation did the MeDIP-Seq data analysis. The single-nucleotide methylation level was estimated based on the mapped MeDIP-Seq reads (corrected by local CpG density) using MEDIPS tool (2).

Human genomic DNA from lung tissues was isolated from LCM samples at different time points during lung development: Early (1d to 1 month old), Mid A (2 months to 1 years old), Mid B (1 year to 2 years old), Mid C (3 to 4 years old), and Late (8 to 9 years old). DNA was isolated using the ZR Genomic DNA-Tissue MicroPrep kit (Zymo Research) as mentioned above. Methylation analysis was performed using Illumina Infinium MethylationEpic BeadChip Kits, which allow interrogation of >850,000 methylation sites at single-nucleotide resolution. 350 ng of DNA was bisulfite converted using EZ DNA Methylation Kits (Zymo Research) and subsequently processed for HumanMethylationEPIC BeadChips, following manufacturer's instructions. Following hybridization, BeadChips were scanned using the Illumina HiScan System. We used the illumina provided Integrated Analysis Software (GenomeStudio Software), to obtain methylation β values, which represents the absolute methylation level at single-nucleotide resolution. As we did for the mouse methylation data, for each gene, we calculate the mean methylation β values for all probes in the promoter region of the gene and use this value for the iDREM reconstruction model.

Proteomic analysis

All specimens were prepared and analyzed independently without pooling. Proteins were extracted using tissue protein extraction reagent (T-PER, Thermo) as per manufacturer's instructions. Protein concentrations were determined with the EZQ protein assay (Life Sciences). For each sample, two micrograms of protein were concentrated using a Savant SpeedVac® Concentrator (Thermo) and run part way (short stack) on a 10% SDS Bis-Tris gel (Life Sciences). The proteins in gel pieces were digested, followed by acidification and extraction, then analyzed by LCMS. One microgram (10μl) of peptide digests was analyzed for each sample in duplicate, injected using a chilled Surveyor Auto Sampler onto a Surveyor HPLC plus

(Thermo Scientific, San Jose CA). This system was run in-line with a Thermo Orbitrap Velos Pro hybrid mass spectrometer equipped with a nano-electrospray source (Thermo Scientific, San Jose CA). All data was collected in CID mode. Following each parent ion scan with mass range of 350-1200*m/z*, fragmentation data was collected on the top-most intense 15 ions. The XCalibur RAW files were centroided and converted to MzXML and the mgf files were created using both ReAdW and MzXML2Search respectively (http://sourceforge.net/projects/sashimi/). The data were searched using the latest version of Mascot Distiller. Searches were performed with a reverse and forward concatenated human subset of the UniRef100 database, which includes common contaminants such as digestion enzymes and human keratins. The final analysis of LCMS generated data was carried out using Expressionist software from Genedata. This package assigns high confidence protein ID's to each ion and aligns mass and time tags from ion plots generated from the post LCMS run. This was then followed by exporting of *.txt matrix files for external statistical analysis, or for internal custom statistical analysis carried out in the seamless module Analyst. Identified peptides were merged with the quantitative data after being filtered and grouped using Peptide and Protein Prophet through the Trans Proteomic Pipeline (TPP; Seattle Proteome Center, Seattle, WA). Only peptides with charge state of ≥ 2+, a minimum peptide length of 6 amino acids, were accepted for analysis. Following the modeling of the entire dataset, the peptide and protein probabilities was set at >80% and >99% respectively, with FDR's based on reverse concatenated searches set at a cut-off of <1%.

Quality of RNA and protein obtained by laser capture

In general, the quality of RNA and protein was excellent. For RNA, 3 LCM caps were pooled per mouse, with multiple mice at each of 15 time points. The average yield of RNA was 800 ng per sample (minimum 300 ng, max 1000 ng). The RNA Integrity Number (RIN) by Agilent Bioanalyzer was used to determine RNA quality, and 100 ng of RNA was then used for RNA sequencing (Illumina TruSeq Stranded with Ribosomal removal). The average mapping rate was 89% with average mapped reads of 18,873,503. Data processing steps included quality control and checking for adapter contamination rate and trimming, followed by mapping by STAR (to map rads back to genome), then calculating FPKM using cufflinks, and finally linear regression to assess the p value of FPKM changes over time. For protein, each sample (one sample consisting of 3 caps per specimen, for a total of 45 specimens) was lysed in 30 microL of T-PER lysis buffer supplemented with protease and phosphatase inhibitors (10 microL per cap), then loaded on a 10% Bis-Tris gel, then stained overnight with colloidal Coomassie. Each lane was excised and digested in-gel with trypsin overnight. The staining indicated good protein concentrations without smearing. Only proteins at >99% CI, with <1% FDR, and > or = 2 peptides were considered for analysis. While good quality of tissue by LCM is not easy to obtain, in this study, the good quality of tissue could be attributed to rapid collection and freezing of mouse lung in OCT, followed by appropriate validated protocols for LCM. RNA and protein quality were also evaluated using well established Standard Operating Protocols (SOP). These protocols are all available for public download on the LungMAP.net website (https://lungmap.net/resources/sop-search-page/).

IRF-1 In-Situ hybridization experiment

For the IRF-1 In-Situ hybridization experiment, lung tissue specimens were isolated from C57BL/6 WT mice at selected time points, fixed 12-16 hours in normal buffered formalin (10 % NBF), embedded in paraffin and freshly cut on super frost plus slides and kept in 4 degree C until performing ISH. In Situ Hybridization was performed using RNAscope Multiplex Fluorescent V2 assay (Advanced Cell Diagnostics, Newark, CA). Briefly, after baking slides in 60° for one hour and deparaffinization, pretreatment steps were applied based on manufacturer's protocol. RNASope 4-Plex Ancillary kit for multiplex fluorescent reagent kit v2 (Cat. No. 3231201) were used for hybridization and amplification with mouse specific RNA probes for IRF-1 combined with Opal Fluorophore 520 from Perkin Elmer, according to both manufacturers' protocols. At the final stage, DAPI was applied for nuclear staining. Vectra Polaris Automated Quantitative Pathology Imaging System from Perkin Elmer was used for imaging. InForm® image analysis software from Perkin Elmer and Image J software were used for analysis and quantification.

Data integration and modeling

Selection of time points to profile

Past work on the analysis and modeling and development and response processes, including our own prior work on lung development (3) has relied primarily on ad-hoc methods to determine which time points to profile. This can lead to several problems when trying to combine different datasets for modeling a specific biological process. First, each dataset on its own may be sampled in a way that misses key events (if they occur between two consecutive sampling points). In addition, integrating different data types, when each is sampled at different time points, is challenging. Indeed, prior work on lung development differed substantially in the set of time points that were sampled, even when profiling the same type of data. To address this issue, we have recently developed the Time Point Selection Method (TPM) method which can be applied to determine the appropriate times to sample given the number of points that can be profiled (which are often limited due to budget or sample availability constraints) (3). TPM relies on a densely sampled subset of genes relevant to the process being studied and uses an iterative greedy algorithm to perform a combinatorial optimization search for selecting the most appropriate time points out of the original sampled ones. In this study we used TPM to select the set of points in which we performed the RNA-Seq and miRNA experiments, and to select a subset of these points for the methylation and proteomics.

The Dynamic Regulatory Events Miner (DREM)

DREM (4-6) integrates static, general, TF-DNA binding interactions data and condition specific (in our case lung development) time-series gene and miRNA expression data to determine the set of TFs and miRNAs that control gene expression over time. The basic idea behind DREM is to try and use the static interaction data (protein-DNA and miRNA-RNA) to explain why a set of co-expressed genes start to diverge at a specific point in time. This allows the method to identify both the TFs and miRNAs controlling these divergence events and to associate a specific time with these (initially static) interactions. See Figure 1 for an example of such assignments. To identify these bifurcation points and the set of TFs and miRNAs that control them, DREM combines a machine learning method termed Input Output Hidden Markov Model (IOHMM) with a logistic regression classifier. DREM has been applied to study several developmental and response processes including lung development (5, 7, 8).

Integrating proteomics and protein-protein interactions

Most work on the reconstruction of dynamic regulatory networks has focused on the use of RNA-Seq and to a lesser extent miRNA data. While some methods further integrated these datasets with proteomics data (for example, SDREM (9)) to date these have been static interaction datasets. In contrast, in our study we have also profiled a complementary time series proteomics data set that we integrated into the DREM model. While there is conflicting evidence about the correlation between mRNA and protein levels, for TFs specifically several studies indicate that expression levels are usually not enough to determine their activity levels. We have thus used the proteomics data to improve our ability to detect the time of TF activation. Specifically, we look for two lines of evidence to determine such activity. The first is the level of the protein itself and the second is the likelihood of a post-translational interaction or modification that leads to its activation. For the former we use the proteomics data directly. For the latter we combine protein-interaction data with the proteomics data as follows: We look at the average protein levels of its interacting partners and if these partners are expressed at a high level we increase our belief in the activity of the TF, even if the TF itself is not over expressed (Figure 1 A). Specifically, we set the activity level of a TF at each time point to:

$$\mathrm{ATF}_{Raw}(x,t) = \frac{1}{|Y|}\sum_{y\in|Y|} P(x,t)P(y,t)PPI(x,y)$$

$$\mathrm{ATF}(\mathrm{x},\mathrm{t}) = \frac{1-e^{-w_{ATF}ATF_{Raw}(x,t)}}{1+e^{-w_{ATF}ATF_{Raw}(x,t)}}$$

Where ATF(x,t) represents the inferred activity of TF x at time t based on the proteomics data, Y denotes the set of proteins that interact with x and PPI(x; y) is the static protein-protein interaction strength between TF x and y obtained from STRING v10 database (10). P(x, t), P(y, t) represent the protein level for TF x and interacting protein y respectively. As we discuss below, this activity is then used as a dynamic prior by our model in order to better determine which TFs regulate which bifurcation event.

Integrating epigenetics data

While we cannot usually obtain direct time series measurements for the binding of TFs (for example, time series ChIP-Seq data for all TFs, which requires several additional experiments for each time points) we can often obtain global indirect information about such events. For example, methylation data was shown to correlate with other epigenetic datasets (11-13) and with binding for several TFs (14-16) and can be profiled globally. Here we use DNA methylation data to obtain a prior on the dynamic binding events for different TFs. DNA methylation is often thought to prevent TF binding by changing the chromatin structure which restricts access of TFs to promoter regions (17). We thus use the time series methylation data to identify "silenced" TFs. These are TFs that, while active, may seem to be inactive for some targets because their binding sites for these targets are methylated. In the original DREM method such TFs would be assigned a low score (since several of their targets are inactive) and would thus may be wrongly removed from the model. For this, we revise the static prior interaction map used by DREM (that assigns each TF-target pair a likelihood of being a target) and reduce this likelihood for genes with methylated promotors. This reduction places more weight on non-methylated targets when compared to methylated ones and so may allow a better identification of active TFs. Specifically, we use the following as the methylation score:

$$\mathrm{Mr(y, t) = 1 - methyl(y, t)}$$

Where Mr(y, t) represents the regulation "score" for gene y at time t based on the given methylation data. methyl (y, t) is the average methylation of the promoter of gene y at time point t. We next discuss how we combine this score with the other interaction and activity values to infer a comprehensive regulatory model.

Inferring dynamic "input": TF-DNA regulation map

We now explain how the different dynamic data sources are combined to derive a dynamic prior for the regulation of a gene by a TF. This dynamic prior is then combined with the dynamic gene expression information of the target gene in order to group genes in paths and infer TF activity. For the dynamic prior we combined the TF activity value *(ATF(x, t))* derived from the time series proteomics data, the methylation score *MR(y, t)* and a static TF-DNA prior $R_{static}(x, y)$ (Figure 1 B). The way these are combined is as follows:

$$\mathrm{R(x, y, t)} = \mathrm{R_d}(x, y, t)R_v(x, y, t)$$

$$\mathrm{R_v}(x, y, t) = \frac{1 - e^{-w_R|\Delta ATF(x,t)|TBS(x,y,t)}}{1 + e^{-w_R|\Delta ATF(x,t)|TBS(x,y,t)}}$$

$$\mathrm{R_d}(x, y, t) = \begin{cases} -1, & if \Delta ATF(x,t)\Delta Ex(y,t) < 0 \\ 1, & if \Delta ATF(x,t)\Delta Ex(y,t) > 0 \\ 0, & else \end{cases}$$

$$\mathrm{\Delta ATF(x, t) = ATF(x, t) - ATF(x, t - 1)}$$

$$\mathrm{\Delta Ex(y, t) = E\ x(y, t) - Ex(y, t - 1)}$$

$$\mathrm{TBS(x, y, t) = Mr(y, t)R_{static}}(x, y)$$

Where *R(x, y, t)* is the dynamic prior for the regulation by TF *x* of target gene *y* at time point *t*. $R_d(x, y, t)$ represents the Regulation direction (activation or repression) and $R_v(x, y, t)$ represents the regulation strength between TF *x* and target gene *y* at time *t*. *Ex(y, t)* represents the expression of *y* at time point *t*. *ATF(x, t)* denotes the estimated TF 'activity' based on proteomics and PPI data. $R_{static}(x, y)$ denotes the strength of binding of TF *x* on gene *y* based the static TF-DNA regulation used in the study. *Mr(y, t)* represents the TF regulation `score' for gene *y* at time point *t* based on the methylation data. Basically, we first combine the two terms that are based on DNA studies (in the TBS terms which combines the methylation and ChIP-Seq or sequence information) and then combine that with the activity value. The value derived for *R(x, y, t)* is then used by DREM (Figure 1 D) as we discuss below.

Likelihood function optimized by the IOHMM

The TF dynamic prior is used to create a regulatory matrix for each time point where the rows represent different TFs and the columns represent genes. Each entry in that matrix is the computed *R(x, y, t)* for that time point. In addition, we also obtain a dynamic prior for the regulation of genes by miRNAs, denoted $R_m(x, y, t)$ as described before (6). Let Og = (Og(1),Og(2), …,Og(n - 1) be the log ratio expression values for gene *g* at time points 1 to *n* -1 (ratios are w.r.t. time point 0). Let *h* represent a hidden state in the model, and assume that gene *g* is assigned to state h_a at time *t* - 1. Then the probability that it would transition to state h_b at *t* is defined as $P(H_t = h_b|H_{t-1} = h_a, I_{g,t})$, where $I_{g,t}$ represents an input vector for gene *g* at time *t* from the dynamic regulatory matrix. This probability is set to 0 if h_b is not a child of h_a and 1 if h_b is the only child of h_a. If h_a has two or more descendants, then the transitions are probabilistic. To learn the parameters for these transitions (which imply which TFs are active at each transition point) we use a logistic regressor as discussed previously (4). The actual learning part involves a search over model and parameters which maximize a global likelihood function (5).

Interactive Visualization

Prior models of dynamic networks usually plotted the expression levels and regulators as a graph (similar to Figure 2 upper part). While this provides useful information on the grouping of genes and the TFs that regulate them, it does not permit interactive exploration and makes it hard to further add other types of data to the model. We have thus also extended the visualization capabilities of DREM by combining it with an interactive display (Figure 2 bottom part). The interactive visualizations were enabled by the iDREM software. Please refer to the iDREM Github page (https://github.com/phoenixding/idrem#interactive-visualization) for detailed descriptions of the interactive visualization.

Comparison to human models

Grouping of human samples

We combined the following four sets of human samples: Birth (Day 1, Day 5), Early infancy (24 Days, 2 Months, 3 Months), Toddler (7 Months, 19 Months, 20 Months, 21 Months, 3 Years, 4

Years), School age (8 Years, 9 Years). Subsequent analysis used the median of each gene for each of these sets.

Aligning mouse and human data

We first extracted top human genes with mouse orthologs based on Ensembl (18). Then we selected top 5000 genes with largest expression correlation between human and mouse. We used splines to represent continuous gene expression for each dataset and tested a number of continuous alignment methods for mapping the human time points to mouse time points. We employed a line search algorithm to search for the optimal parameters for each function we tested (Supporting Methods). We next iterated for each function as follows. We first weight all genes equally. Next, given the optimal mapping so far, we calculate the Pearson correlation for the selected orthologous genes and update the weight for each gene based on its agreement with the alignment by setting: $w_{update}(g) = w_{current}(g) + pearsonr(g)$ and normalizing such that $\sum_{g \in All} w_{update}(g) = 1$. Using this strategy, genes with larger correlation (good mapping with current mapping) that are likely related to the process being studied obtain higher weights while the weight for genes that are less involved in the process is reduced. The best function we found was an exponential alignment $y = 2.9\log(x) + 5.1$ where y is the mouse day and x is the human day. We also calculated the p-value for the obtained mapping by permuting the ordering of the human and mouse data. For each permutation, we used the same analysis procedure as discussed above and obtained a p-value=0.029. Note that since the number of time points is limited, some of the permutations are actually very close to the correct ordering itself which is why the p-value cannot go much lower.

Reconstructing human developmental model

We used iDREM to integrate the different human datasets we collected in order to reconstruct human lung developmental networks. We combined the human data (time series mRNA expression data, time series proteomics data, time series DNA methylation data and static TF-gene and PPI human interaction data) in the same way as we did in mice.

Supporting results

The comparison between models with/without using proteomic and epigenetic datasets

There are 131 Transcription factors predicted (cutoff: top 10 TFs for each node) using the epigenetics and proteomics datasets, whereas only 113 Transcription factors were found without using the epigenetics and proteomics dataset. There are 77 TFs (77/113=68%) shared between these two different models. There are 54 TFs found only with epigenetics and proteomics data, and 36 TFs only found without epigenetics and proteomics data. These 54 TFs with epigenetics and proteomics data are much more significantly enriched in developmental process GO term (6.27e-8) compared with the 36 TFs without epigenetics and proteomics data (5.1e-4). Also, many known lung regulating factors such as RUNX2, CASR, TREM13, and TPX5 can only found with using epigenetics and proteomics datasets as we discussed in the manuscript. The above analyses demonstrated that more supported transcription factors can be identified with using the epigenetics and proteomics datasets.

Supplemental tables

Supplemental Table S1 GSEA enrichment analysis (C2: curated gene sets) of the top 50 regulators for each path

Top 50 Regulators	Top 5 enriched terms

Path A
36 TFs
14 miRNAs
	REACTOME_INTERFERON_GAMMA_SIGNALING (3.18e-15);
	REACTOME_INTERFERON_ALPHA_BETA_SIGNALING(3.18e-15);
	REACTOME_CYTOKINE_SIGNALING_IN_IMMUNE_NE_SYSTEM(2.81e-13);
	REACTOME_INTERFERON_SIGNALING(7.79e-12); ISHIKAWA_STING_SIGNALING(1.37e-11);

Path B
47 TFs
3miRNAs
	PID_SMAD2_3NUCLEAR_PATHWAY(1.41e-19); PID_REG_GR_PATHWAY(2.73e-15);
	PID_AP1_PATHWAY(6.56e-14);
	REACTOME_NUCLEAR_RECEPTOR_TRANSCRIPTION_PATHWAY(6.79e-11);
	TENENDINI_MEGAKARYOCYTE_MARKERS(6.79e-11);

Path C
27 TFs
23miRNAs
	REACTOME_INTERFERON_ALPHA_BETA_SIGNALING(1.55e-21);
	REACTOME_INTERFERON_GAMMA_SIGNALING(3.39e-19);
	REACTOME_INTERFERON_SIGNALING(1.94e-17);
	REACTOME_CYTOKINE_SIGNALING_IN_IMMUNE_SYSTEM(4.41e-17);
	BIOCARTA_EGF_PATHWAY(3.42e-14);

Path D
46 TFs
4 miRNAs
	PID_SMAD2_3NUCLEAR_PATHWAY(2.28e-17);
	KEGG_PATHWAYS_IN_CANCER(4.43e-16);
	PID_REG_GR_PATHWAY(2.38e-13);
	REACTOME_NUCLEAR_RECEPTOR_TRANSCRIPTION_PATHWAY(3.32e-13);
	REACTOME_INTERFERON_ALPHA_BETA_SIGNALING(2.58e-13);

Path E
50 TFs
0 miRNAs
	PID_SMAD2_3NUCLEAR_PATHWAY(1.47e-21);
	PID_AP1_PATHWAY(4.84e-18);
	KEGG_PATHWAYS_IN_CANCER(1.09e-15);
	PID_HIF1_TFPATHWAY(4.06e-14);
	TENEDINI_MEGAKARYOCYTE_MARKERS(4.06-e14);

Path F
26 TFs
	REACTOME_NUCLEAR_RECEPTOR_TRANSCRIPTION_PATHWAY(9.52e-15);
	REACTOME_GENERIC_TRANSCRIPTION_PATHWAY(5.09e-8);

24
miRNAs
	REACTOME_PPARA_ACTIVATES_GENE_EXPRESSION(5.25e-8);
	PID_MYC_REPRESS_PATHWAY(3.87e-7);
	REACTOME_FATTY_ACID_TRIACYLGLYCEROL_AND_KETONE_BODY_METABOLISM(5.77e-7);

Path G
30TFs
20
miRNAs
	PID_AP1_PATHWAY(2.42e-15)
	NAGASHIMA_EGF_SIGNALING_UP(2.14e-11);
	CHASSOT_SKIN_WOUND(3.97e-11);
	PID_SMAD2_3NUCLEAR_PATHWAY(1.13e-10);
	KEGG_MAPK_SIGNALING_PATHWAY(1.13-e10);

Path H
12TFs
38
miRNAs
	COLLIS_PRKDC_SUBSTRATES(4.37e-8);
	PID_SMAD2_3NUCLEAR_PATHWAY(4.37e-8);
	BIOCARTA_MAPK_PATHWAY(4.37e-8);
	BIOCARTA_P38MAPK_PATHWAY(4.16e-7);
	YEMELYANOV_GR_TARGETS_DN(2.02e-6);

Path I 0 TFs 50 miRNAs	NA

Path J
7TFs
43
miRNAs
	REACTOME_GO_AND_EARLY_G1(3.79e-5);
	REACTOME_G1_PHASE(6.93e-5);
	PID_RB_1PATHWAY(2.39e-4);
	PID_E2F_PATHWAY(2.4e-4);
	RIZ_ERYTHROID_DIFFERENTIATION(2.4e-4);

Path K
15 TFs
35
miRNAs
	REACTOME_G1_PHASE(8e-12);
	PID_RB_1PATHWAY(1.19e-10);
	PID_E2F_PATHWAY(1.78e-10);
	KEGG_CELL_CYCLE(3.86e-9);
	PID_P53_DOWNSTREAM_PATHWAY(3.9e-9);

Path L
50TFs
0 miRNAs
	BENPORATH_SUZ12_TARGETS(3.12e-12);
	PID_E2F_PATHWAY(3.16e-11);
	BENPORATH_EED_TARGETS(3.16e-11);
	BENPORATH_ES_WITH_H3K27ME3(5.21e-11);
	BENPORATH_PRC2_TARGETS(5.49e-9);

Supplemental Table S2 Top GO terms associated with the genes in each path

Path	**Top 5 GO terms (with p-values)**
A	extracellular region part (2.3e-23), defense response (2.43e-20), immune response (1.79e-18), Vesicle (2.25e-18), response to external stimulus (1.45e-17)
B	positive regulation of biological process (4.04e-38), regulation of response to stimulus (1.18e-37), regulation of immune system process (1.72e-32), positive regulation of cellular process (1.09e-31), regulation of multicellular organismal process (2.24e-30)
C	defense response (3.61e-15), immune response (3.14e-23), regulation of immune system process (2.97e-18), cell activation (5.82e-19), leukocyte activation (5.74e-17)
D	defense response (3.73e-12), immune response (2.98e-16), innate immune response (1.99e-11), response to cytokine (7.76e-11), response to interferon-beta (1.56e-10)
E	intracellular (1.23e-51), intracellular part (2.36e-48), cytoplasm (5.16e-47), protein binding (1.60e-36) and single-organism cellular process (1.14e-35)
F	defense response (4.83e-20), immune response (3.23e-16), inflammatory response (4.64e-15), positive regulation of immune system process (7.35e-14) immune effector process (3.52e-13)

G	axoneme assembly (2.407e-8), microtubule bundle formation (3.25e-6), cilium movement (6.56e-5), axonemal dynein complex assembly (7.53e-5), axoneme part (1.45e-4).
H	intracellular (2.30e-49), intracellular part (3.26e-44), cytoplasm (1.14e-40), membrane-bounded organelle (1.98e-41), cytoplasmic part (3.34e-41)
I	intracellular (5.86e-108), intracellular part (6.62e-94), membrane-bounded organelle (1.72e-87), intracellular organelle (1.26e-86), intracellular organelle part (1.39e-82)
J	intracellular (2.97e-49), intracellular part (3.94e-57), membrane-bounded organelle (3.14e-54), intracellular membrane-bounded organelle (6.40e-54), intracellular organelle (7.62e-55)
K	mitotic cell cycle process (9.93e-52), mitotic cell cycle (1.38e-50), cell cycle process (4.05e-50), cell cycle (8.54e-46) and nuclear division (2.38e-45)
L	DNA packaging complex (1.06e-50), nucleosome (9.65e-50), nuclear nucleosome (1.19e-44), protein-DNA complex (5.37e-43), DNA packaging (2.75e-42)

Supplemental Table S3 GO terms (with corrected p-value <1E-4) for top 1k genes that best agreed with the inferred mouse-human alignment

GO biological process complete	**corrected P-value**
sensory perception of chemical stimulus (GO:0007606)	1.29E-17
sensory perception of smell (GO:0007608)	1.64E-14
positive regulation of cellular process (GO:0048522)	4.64E-13
positive regulation of biological process (GO:0048518)	5.15E-13
cellular process (GO:0009987)	9.64E-12
regulation of cell proliferation (GO:0042127)	7.21E-11
positive regulation of cell proliferation (GO:0008284)	2.90E-10
sensory perception (GO:0007600)	3.70E-10
circulatory system development (GO:0072359)	5.82E-10
anatomical structure development (GO:0048856)	8.57E-10
regulation of response to stimulus (GO:0048583)	1.42E-09
multicellular organism development (GO:0007275)	1.52E-09
developmental process (GO:0032502)	3.27E-09
negative regulation of biological process (GO:0048519)	6.53E-09
regulation of multicellular organismal development (GO:2000026)	1.07E-08
regulation of cell adhesion (GO:0030155)	1.48E-08
biological process (GO:0008150)	1.73E-08
positive regulation of multicellular organismal process (GO:0051240)	2.46E-08
positive regulation of developmental process (GO:0051094)	1.14E-07
negative regulation of cellular process (GO:0048523)	1.32E-07
positive regulation of cell adhesion (GO:0045785)	1.81E-07
regulation of developmental process (GO:0050793)	2.12E-07
system development (GO:0048731)	2.57E-07
biological regulation (GO:0065007)	7.15E-07
animal organ development (GO:0048513)	1.47E-06

regulation of biological process (GO:0050789)	9.52E-06
negative regulation of multicellular organismal process (GO:0051241)	1.49E-05
positive regulation of metabolic process (GO:0009893)	1.71E-05
regulation of signal transduction (GO:0009966)	1.92E-05
G-protein coupled receptor signaling pathway (GO:0007186)	2.04E-05
anatomical structure morphogenesis (GO:0009653)	2.10E-05
cardiovascular system development (GO:0072358)	3.28E-05
vasculature development (GO:0001944)	3.34E-05
nervous system process (GO:0050877)	3.63E-05
positive regulation of macromolecule metabolic process (GO:0010604)	4.21E-05
regulation of cell differentiation (GO:0045595)	5.40E-05
positive regulation of cell migration (GO:0030335)	5.52E-05
negative regulation of response to stimulus (GO:0048585)	5.53E-05
regulation of cellular component movement (GO:0051270)	5.63E-05
regulation of cell migration (GO:0030334)	6.45E-05
positive regulation of cellular component movement (GO:0051272)	7.17E-05
positive regulation of cell differentiation (GO:0045597)	7.18E-05
positive regulation of cell motility (GO:2000147)	7.23E-05
heart development (GO:0007507)	7.39E-05
regulation of intracellular signal transduction (GO:1902531)	7.76E-05
positive regulation of locomotion (GO:0040017)	8.58E-05

Supplemental Table S4 Examples of regulators for lung development and associated processes in paths of the computational model

Path	**Known regulator(s) for lung development in Paths (key regulators selected from Sankey diagram)** ˋ	**Processes in lung development (PMID of citations)**
A	AR, RELA	AR seen in nucleus of epithelial cells (PMID 19576262); RELA regulates maturation of AT1 and AT2 (PMID 18161062)
B	NFKB1, CREB1, mmu-miR-326, mmu-miR-124	NFKB regulates maturation of AT1 and AT2 (PMID 18161062); CREB1 necessary during later stages of lung development for epithelial differentiation (PMID 2140814); miR-326 is downstream of Shh [Shh is secreted from epithelia and acts on adjacent mesenchyme via Smo and Gli], and regulates Smo and Gli2 (PMID 24617895); miR-124 regulates lung epithelial cell maturation at later lung development, and inhibits NFKB signaling (PMID 26071557)
C	mmu-miR-539	We inhibited miR-539 in the developing murine lung and observed that alveolar development was reduced significantly.
D	PGR, PPARG, RUNX2	Progesterone receptor (PGR) in epithelia important in postnatal lung development, dependent upon glucocorticoid/GR signaling (PMID 17003286);

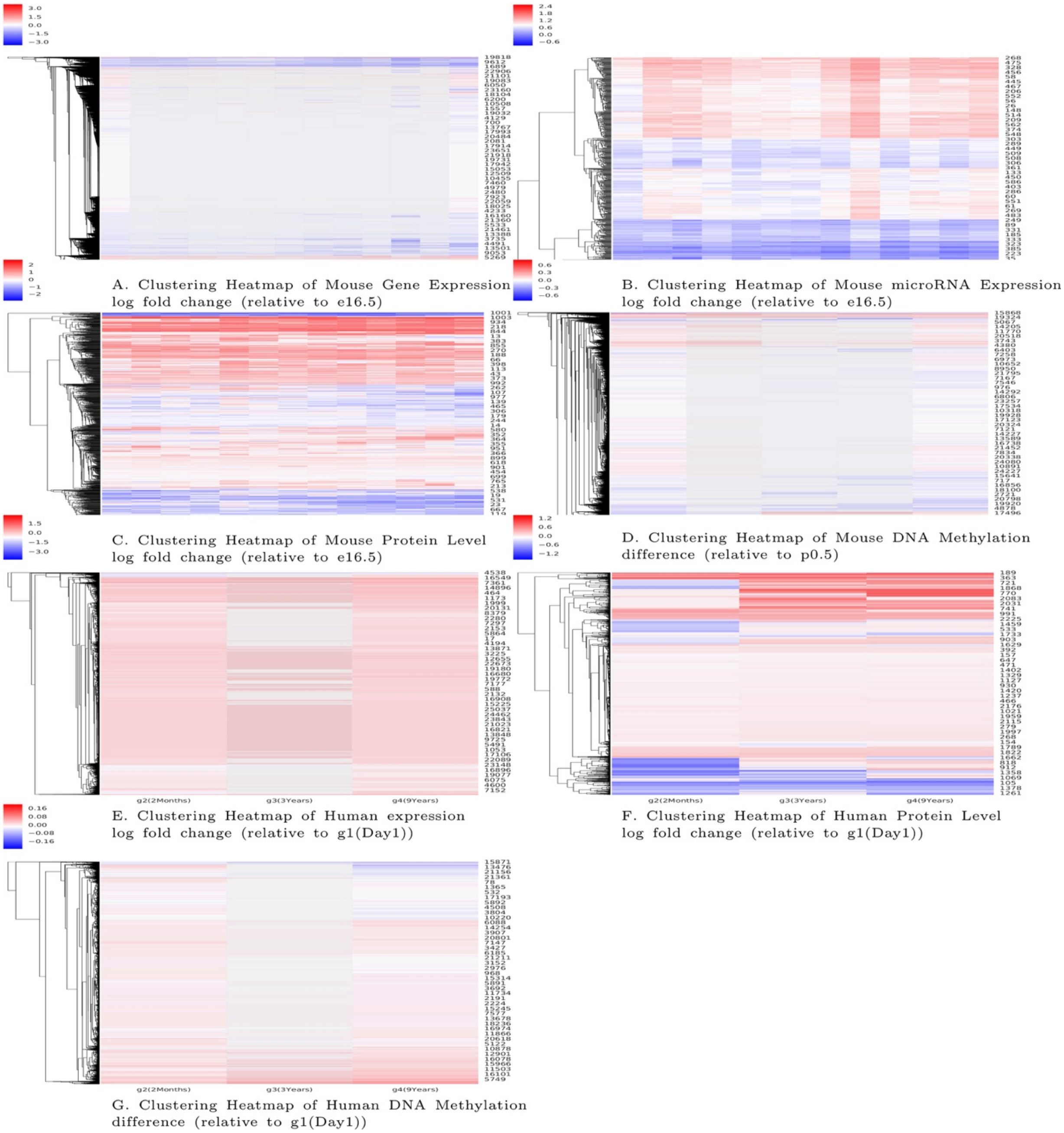
		PPARG in multiple cell types (epithelial, fibroblast, immune) important for lung maturation (PMID 19262292; 16720732); RUNX2 regulates goblet cell differentiation (PMID 27825108), and mediates SPP1 that is a determinant of alveolar development (PMID 24816281)
E	CEBPA, SMAD3, CASR	CEBPA required for epithelial cell differentiation (PMID 22411169). Smad3 is a major downstream signal transducer of TGF-beta pathway, and retarded lung development is seen in Smad3 null mice (PMID 15591413); The extracellular calcium-sensing receptor CASR regulates fluid secretion in the developing lung via CFTR (PMID 26911344)
F	CUX1, mmu-miR-124	CUX1 is necessary for alveolar development – CUX1 knock-out mice died shortly after birth from respiratory failure due to a delay in lung development (arrest in early saccular stage) (PMID 11544187); miR-124 known to regulate proliferative, migratory, and inflammatory phenotype of pulmonary vascular fibroblasts (PMID 24122720)
G	ARNT, NFE2L1, RFX1, JUND, FOS, FOSB, JUNB, TRIM13, mmu-let-7c,mmu-let-7e,mmu-let-7d,mmu-let-7f,mmu-let-7a	The arylhydrocarbon-receptor nuclear translocator (ARNT) is essential for blood vessel development, and arnt -/- embryos are not viable beyond E10.5 (PMID 9121557). NFE2L1 mice are also embryonic lethal (PMID 12968018). Regulatory Factor X (RFX)1 regulates ciliary function (PMID 27451412) and gene expression (PMID 20690903). Jun and Fos combine to form AP-1 transcription factor, that regulates TGF-beta signaling, important for lung alveolar development (PMID 18321928). Let-7 microRNA is regulated by FGF and in turn regulates TGF-beta (PMID 23200853)
H	USF1	Upstream stimulatory factor-1 (USF1) is involved in developmental and hormonal regulation of SP-A gene in AT2 cells (PMID 12576297, 9287355).
I	mmu-miR-495	miR-495 is an inducer of SOX-9 (well known in lung development), and is involved in development of stem cells and cartilage as well as other biological processes (cell proliferation, invasion, apoptosis (PMID 28454357).
J	TBX5	TBX5 necessary for normal lung development at multiple time points (PMID 22876201)
K	RB1, E2F5, E2F2, E2F4	E2F sites are present in the promoter of many genes whose products are involved in cell proliferation. The HDAC1/2-BMP4/RB1 pathway regulates Sox2+ endoderm progenitors in the lung endoderm (PMID 23449471)
L	HMGA2	HMGA2 is required for canonical WNT signaling and regulates cell proliferation during lung development (PMID 24661562)

Supplemental Table S5 Verified regulatory relationship in the existing literature observed in the model

Interaction	**Observed in model**	**Existing literature**
miRNA-mRNA	miR-145 regulates Path H (in addition to	miR-145 regulates myofibroblast

	connections between nodes 6 &16, and 3 and 11). ACTA2 and PDGFRA expression is along Path H, which is consistent with myofibroblasts.	differentiation by promoting activation of TGF-beta1 and regulating smooth muscle actin-alpha (ACTA2) (PMID 23457217)
DNA methylation - mRNA	Nrn1 is low from E16.5-P13.5, and increases from P15 to P28. DNA methylation of Nrn1 is high until P10, then lower at P19 and P28. Lrp2 is low until P19, then increases at P23 and P28. DNA methylation of Lrp2 is lowest at P19 and P28. MMP3 increases after P10, and its DNA methylation is highest at P7, and minimal at P10.	Nrn1, Lrp2, and MMP3 are known to be regulated by DNA methylation during lung development (PMID 25387348)
Transcription factor – known targets	NKX2-1 regulates PATH E, and DKK3, CD40, and SCRN2 expressions are along PATH E	DKK3 (p 6.23667E-20), CD40 (p 4.15667E-18), and SCRN2 (p 8.893E-18) are targets of NKX2-1 at E19.5 (PMID 22242187)

Supplemental figures



A. Clustering Heatmap of Mouse Gene Expression log fold change (relative to e16.5)

B. Clustering Heatmap of Mouse microRNA Expression log fold change (relative to e16.5)

C. Clustering Heatmap of Mouse Protein Level log fold change (relative to e16.5)

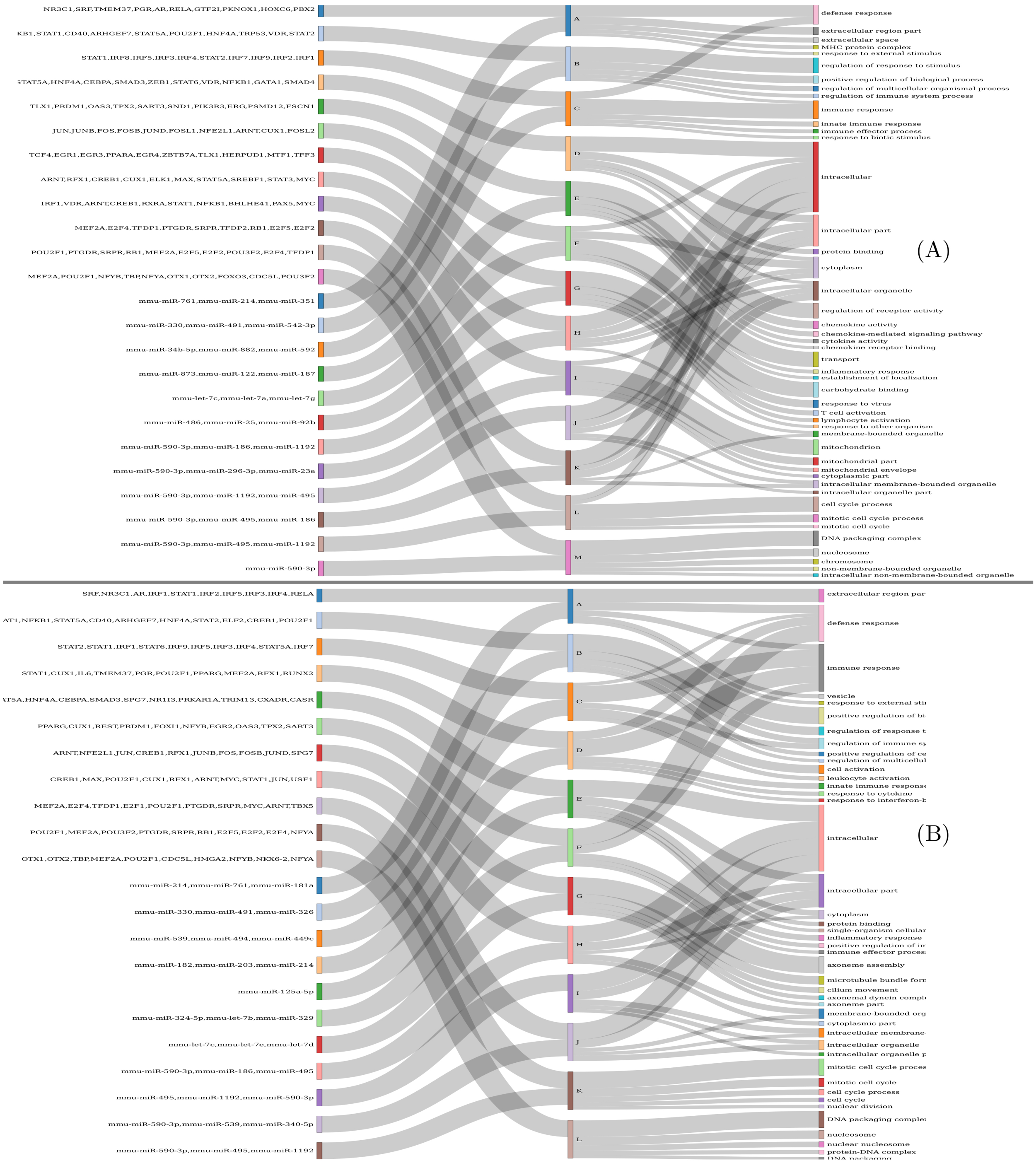
D. Clustering Heatmap of Mouse DNA Methylation difference (relative to p0.5)

E. Clustering Heatmap of Human expression log fold change (relative to g1(Day1))

F. Clustering Heatmap of Human Protein Level log fold change (relative to g1(Day1))

G. Clustering Heatmap of Human DNA Methylation difference (relative to g1(Day1))

Supplemental Figure S1. The clustering heatmaps of mouse and human datasets. A. The clustering heatmap of log2 fold change of the mouse gene expression data. B. The clustering heatmap of log2 fold change of the mouse microRNA expression data. C. The clustering heatmap of log2 fold change of the mouse proteomic data. D. The clustering heatmap of the mouse DNA methylation difference (relative to p0.5). E. The clustering heatmap of log2 fold change of the human gene expression data. F. The clustering heatmap of log2 fold change of the human proteomic data. G. The clustering heatmap. of the human DNA methylation difference.



Supplemental Figure S2. The Sankey Plots for the models with/without using proteomic and epigenetic datasets. The middle columns of the Sankey plots represent the predicted paths, in which the genes have similar expression patterns. The left columns are top regulating factors and the right columns are enriched GO terms associated with each of the paths. (A) The model without using proteomic and epigenetic datasets. (B) The model with using the proteomic and epigenetic datasets. Many known lung development regulators such as RUNX2, CASR, TREM13, and TPX5 can only be identified in the model with using the proteomic and epigenetic datasets.

References

1. Taiwo O*, et al.* (2012) Methylome analysis using MeDIP-seq with low DNA concentrations. *Nature protocols* 7(4):617-636.
2. Lienhard M, Grimm C, Morkel M, Herwig R, & Chavez L (2013) MEDIPS: genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. *Bioinformatics* 30(2):284-286.
3. Kleyman M*, et al.* (2017) Selecting the most appropriate time points to profile in high-throughput studies. *Elife* 6.
4. Ernst J, Vainas O, Harbison CT, Simon I, & Bar-Joseph Z (2007) Reconstructing dynamic regulatory maps. *Molecular systems biology* 3:74.
5. Schulz MH*, et al.* (2012) DREM 2.0: Improved reconstruction of dynamic regulatory networks from time-series expression data. *BMC Syst Biol* 6:104.
6. Schulz MH*, et al.* (2013) Reconstructing dynamic microRNA-regulated interaction networks. *Proc Natl Acad Sci U S A* 110(39):15686-15691.
7. Chang KN*, et al.* (2013) Temporal transcriptional response to ethylene gas drives growth hormone cross-regulation in Arabidopsis. *Elife* 2:e00675.
8. Xiao Y & Segal MR (2009) Identification of yeast transcriptional regulation networks using multivariate random forests. *PLoS computational biology* 5(6):e1000414.
9. Gitter A, Carmi M, Barkai N, & Bar-Joseph Z (2013) Linking the signaling cascades and dynamic regulatory networks controlling stress responses. *Genome research* 23(2):365-376.
10. Szklarczyk D*, et al.* (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic acids research* 43(Database issue):D447-452.
11. Jones PA & Takai D (2001) The role of DNA methylation in mammalian epigenetics. *Science* 293(5532):1068-1070.
12. Johnson L, Cao X, & Jacobsen S (2002) Interplay between two epigenetic marks. DNA methylation and histone H3 lysine 9 methylation. *Current biology : CB* 12(16):1360-1367.
13. Li X*, et al.* (2008) High-resolution mapping of epigenetic modifications of the rice genome uncovers interplay between DNA methylation, histone methylation, and gene expression. *Plant Cell* 20(2):259-276.
14. Zilberman D, Gehring M, Tran RK, Ballinger T, & Henikoff S (2007) Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nat Genet* 39(1):61-69.
15. Watt F & Molloy PL (1988) Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. *Genes & development* 2(9):1136-1143.
16. Tate PH & Bird AP (1993) Effects of DNA methylation on DNA-binding proteins and gene expression. *Curr Opin Genet Dev* 3(2):226-231.
17. Harmston N & Lenhard B (2013) Chromatin and epigenetic features of long-range gene regulation. *Nucleic acids research* 41(15):7185-7199.
18. Aken BL*, et al.* (2016) Ensembl 2017. *Nucleic acids research* 45(D1):D635-D642.