SMU Data Science Review

Volume 3 | Number 2 Article 7

2020

The Transcript Profile Changes With Developmental Maturation of Fetal Lung Type 2 Cells: An Analysis of RNAseq Data

Heber C. Nielsen
Southern Methodist University, hcnielsen@smu.edu

Volodymyr Orlov
Southern Methodist University, vorlov@mail.smu.edu

Rebecca Holsapple
Southern Methodist University, rholsapple@mail.smu.edu

Monnie McGee
Southern Methodist University, mmcgee@smu.edu

The Transcript Profile Changes With Developmental Maturation of Fetal Lung Type 2 Cells: An Analysis of RNAseq Data

Heber C Nielsen[1,2], Volodymyr Orlov[1], Rebecca Holsapple[1], and Monnie McGee[3]

[1] Master of Science in Data Science, Southern Methodist University, Dallas TX 75275 USA {hcnielsen,vorlov,rholsapple}@smu.edu
[2] Department of Pediatrics, Tufts Medical Center, Boston, MA, USA heber.nielsen@tufts.edu
[3] Department of Statistical Science, Southern Methodist University, Dallas, Texas, USA mmcgee@smu.edu

Abstract. In this paper, we utilize next-generation sequencing (NGS) data from the LungMap project [1] to identify and characterize the developmental RNA transcriptome in alveolar epithelial type II (AT2) cells of embryonic mouse lungs of gestational ages embryonic days 16 (E16) and 18 (E18). Late gestation lung cellular maturation is necessary for survival at birth [4]. Using R and the BioConductor packages for RNAseq analysis, we analyze changes in the mouse lung AT2 cell RNA transcriptome as this maturation process takes place. We particularly identify the cluster of genes whose expression changes markedly between immature (E16) and mature (E18) lungs which can be used to define cell pathways that appear critical for the maturation process. Our results show that there are 98 differentially expressed genes with 82 genes where differences in counts cannot be attributed to difference in sample origin. We were surprised to identify substantial differences in RNA expression between two experienced lab groups that use identical protocols.

1 Introduction

Preterm birth, defined as birth before 37 weeks gestational age, is the leading cause of death in the first year of life, and the major cause of death or chronic morbidity following birth [8]. The primary basis of morbidity and mortality following preterm birth is the Respiratory Distress Syndrome (RDS), a common breathing disorder that affects premature newborns shortly after birth. RDS is generally caused by a lack of sufficient amounts of pulmonary surfactant in the distal structures of the lung where gas exchange occurs to transfer oxygen into the blood and remove carbon dioxide. Pulmonary surfactant is a complex mixture of specific phospholipids and proteins that is required to keep the lungs fully expanded at the end of each breath, thus allowing easy, effortless breathing. Its main function is to reduce the surface tension that develops with each expiration at the air/liquid interface in the lung produced by the small amounts of water

that are present in the lung. Without surfactant the surface tension forces cause progressive atelectasis, a collapse of the saccules and alveoli, which are the site of gas exchange, and thus progressively inadequate oxygen intake and carbon dioxide removal.

Most infants with RDS are those born before 32 weeks gestation since maturation of the lung's ability to synthesize surfactant takes place at 32-34 weeks. Surfactant is synthesized in specialized cells in the saccules and alveoli called Type II epithelial cells (AT2 cells). It is known that AT2 cell maturation is critical for surfactant production but the specific mechanisms controlling type II cell maturation are incompletely defined [4].

In this study we are working to advance the understanding of fetal lung maturation and AT2 cell development. The Lung Map Consortium is a multi-institutional group of lung developmental biology investigators funded by the Lung Division of the National Heart, Lung and Blood Institute at the National Institutes of Health that collects, annotates, validates, and stores databases of gene expression, protein expression, metabolite expression and structural imagery of the developing human and mouse lungs for use by interested investigators as source data for lung biology investigations. We utilize Lung Map repository data to identify developmental patterns of AT2 cell gene expression surrounding the timing of lung maturation in the embryonic mouse. Specifically, we make use of Lung Map RNAseq data sets obtained from mouse lungs AT2 cells at embryonic days 16 (E16) and 18 (E18) of mouse gestation. These gestational ages bracket the time at which surfactant synthesis normally surges in the fetal mouse. We use R, in particular the Bioconductor suite of R packages designed for RNAseq analysis [6, 7, 10, 11], to identify the developmental RNA expression signatures at these two developmental ages. We further define how the signatures change, with attention to relative changes between the two gestations in the expression of each gene in the mouse transcriptome. Strong changes in expression are particularly noted for further analysis.

We expect the result of these studies to identify groups of genes whose expression falls or rises with progressing development in patterns that support and extend the knowledge of the regulation of the development of surfactant synthesis. We expect that with this information future studies utilizing pathway analysis will be able to categorize differentially regulated genes according to cellular processes involved in the development of mature AT2 function. The outcome of this project will be novel insight and understanding of fetal lung maturation, that become a basis for new innovative studies of the developmental control of fetal maturation, with the goal of identifying new therapeutic targets for preventing and treating RDS and its complications.

2 Data

Three data sets, each consisting of RNAseq data from a set of experiments utilizing male and female lungs from E16.5 and E18.5 days of gestations, as well as various postnatal days, were used. The lungs were obtained by investigators

within the LungMap Consortium, and prepared for RNA sequencing. Sequencing was performed using a HiSeq 2500 System (Illumina, San Diego, CA). Overall, among the three data sets, 30 female and 20 male samples were sequenced.

The experiments used mice from the inbred laboratory strain C57Bl6, a commonly used laboratory mouse strain widely available. Pregnancies were timed based on the appearance of a vaginal plug the morning after breeding; this was designated day E0.5. On day E16.5 or E18.5 the pregnant females were sacrificed by CO_2 inhalation, individual fetuses obtained and weighed, and the lungs removed under sterile conditions. For preparations of sex-specific samples the lungs were kept separate, otherwise the lungs were pooled. Fetal sex was subsequently identified by genotyping for the presence or absence of the Y chromosome. The lungs were prepared for RNA sequencing by cell dispersion and the alveolar epithelial cell population isolated by cell sorting. cDNA libraries were prepared and used to construct RNAseq libraries using TruSeq DNA Exome library preparation and exome enrichment reagents (Illumina, San Diego CA). RNA sequencing was performed as described by Treutlein, et al [13] using an Illumina HiSeq 2500 sequencing platform. Sequencing was done on 100bp paired end to a depth of 2-5 million reads. Raw data were output to *.fastq files, then pre-processed and sequence alignment performed, then stored in compressed, Binary Sequence Alignment/Map (BAM) formatted data (.BAM) files which were archived on the LungMap repository server.

The sequencing data used in this study were from 2 separate sequencing experiments; the data files are identified on LungMap as LMEX0000001630, hereafter referred to as data set 1630, and LMEX0000000684, hereafter referred to as data set 684. The cell preparations, sequencing, and pre-processing were done at two different institutional members of the LungMap consortium. Data set 1630 consisted of one replicate each of male and female lungs, each of gestational age E16.5 and E18.5. Data set 684 contained data from two replicates each at gestational ages E16.5 and E18.5, all from pooled-sex samples. The attributes of data sets 684 and 1630 are presented in Table 1.

Table 1. Attributes of data sets 684 and 1630. File origin is either University of Alabama (UAB) or Cincinnati Children's Hospital Medical Center (CCHMC). Sex is either Female (F), Male (M) or Both (B).

File ID	Data Set	Library Size	Gestation Age	File Origin	Sex
1	684	26945282	E16.5	CCHMC	B
2	684	21750402	E16.5	CCHMC	B
3	684	25986466	E18.5	CCHMC	B
4	684	25112237	E18.5	CCHMC	B
5	1630	12271916	E16.5	UAB	F
6	1630	15153460	E16.5	UAB	M
7	1630	15329422	E18.5	UAB	F
8	1630	13570271	E18.5	UAB	M

3 Differential expression analysis workflow

The goals of this analysis are to identify the transcriptomic signatures in AT2 cells isolated from E16 and E18 male and female mice, and begin to compare those signatures over time to identify groups of genes that are differentially expressed as AT2 maturation occurs.

Because the analysis uses counts of short sequences of nucleotides the properties of count data must be identified before doing statistical analysis. A major property to evaluate is the possible sources of variability. Three sources of variability were considered: Poisson variability, biological variability and systematic variability. The analysis seeks to identify systematic variability due to the difference between conditions after the other two sources have been controlled for. Biological variability arises from inherent differences between individuals in expression levels. Large differences in the genetic code of specific genes are also possible contributors to biologic variability, but this is unlikely within a population from an inbred strain of mice. Poisson variability arises from the fact that RNA is normally created from DNA in fragments of any individual gene, and in the sequencing process each RNA fragment that is present either is or is not captured. Correct analysis and interpretation of the data requires the assumption that all three sources of variability are present in the data. Another important aspect of data analysis is that while genes are not expressed independently of each other, the analysis makes comparisons one gene at a time, thus correction for multiple comparisons must be done. The pipeline used to analyze the data accounts and adjusts for all of these properties and is displayed in Figure 1.

Our analysis starts with the RSamtools package shown at the beginning of the pipeline. This tool makes it possible to read compressed, Binary Sequence Alignment/Map (BAM) formated data, a format that is frequently used to to store and share genetic sequencing data. RSamtools transforms BAM files into a format that can be understood by the downstream packages. The raw sequence data contained in the BAM files were mapped to a reference mouse genome database mm10, from the University of California Santa Cruz Genomics Institute Genome Browser (https://genome.ucsc.edu/cgi-bin/hgGateway).

3.1 Feature counts and data normalization

Sequencing output provides raw counts. It is important to not use raw counts for differential expression and related analyses since the depth of sequencing results in libraries with higher counts. Therefore, raw counts were transformed into a log2- counts per million (1CPM) scale that helps to normalize for library size differences. After transformation to 1CPM, genes that showed extremely low expression were removed, since genes that are not expressed or expressed at a low level add distributional bias without providing any additional information to the results. Removing these genes also has the advantage of reducing the overall number of statistical tests that should be performed. Figure 2 shows the density function of the 1CPM counts for raw and filtered data.

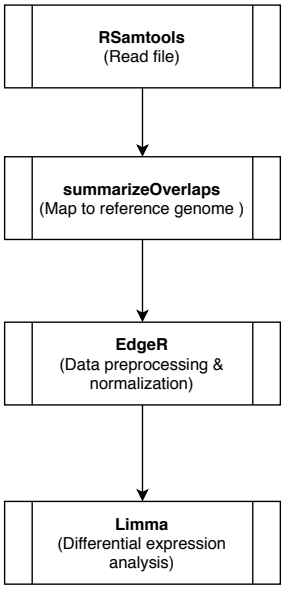

Fig. 1. In this work we use sequential process that consist of 4 steps: data transformation, mapping to reference genome, normalization, and differential expression analysis.

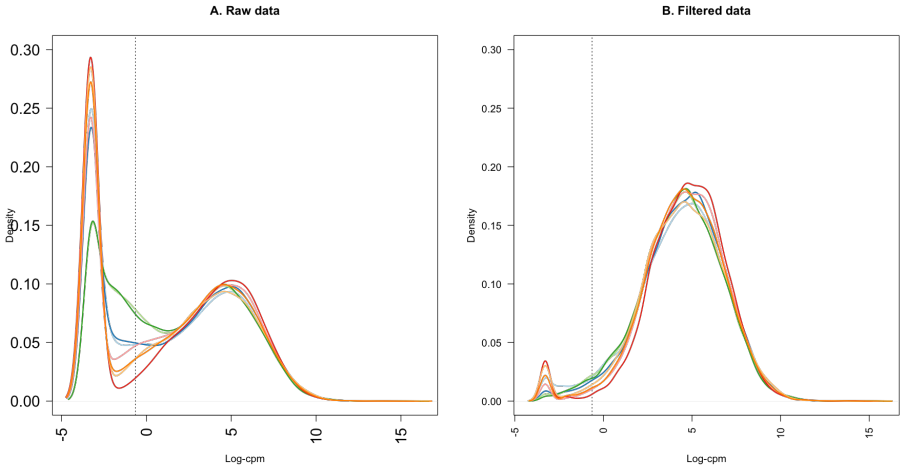

Fig. 2. The density function of the 1CPM counts of raw and filtered data. Left-hand side shows density of 1CPM counts before filtering while right-hand side shows the same data where genes with low expression are removed.

The feature counts were normalized using the method of Trimmed Mean of M-values [12] (TMM). The choice of normalization method has great impact on results and is crucial for correct analysis of RNAseq data [2]. The numerical value of genes can vary due to biological and technical variability. Normalization helps to correctly infer differential signals in RNAseq data and to remove the distributional influence of external factors that are not of biological interest. The TMM method works under the assumption that the majority of genes are not differentially expressed. It finds non-differentially expressed genes that have the same level of expression and uses these genes to calculate normalization factors that are applied to each sample to correct for non-biological variability. Factors are calculated using adjusted mean, a method of averaging that removes a small designated percentage of the largest and smallest values before calculating the mean. Figure 3 illustrates the results of the data normalization.

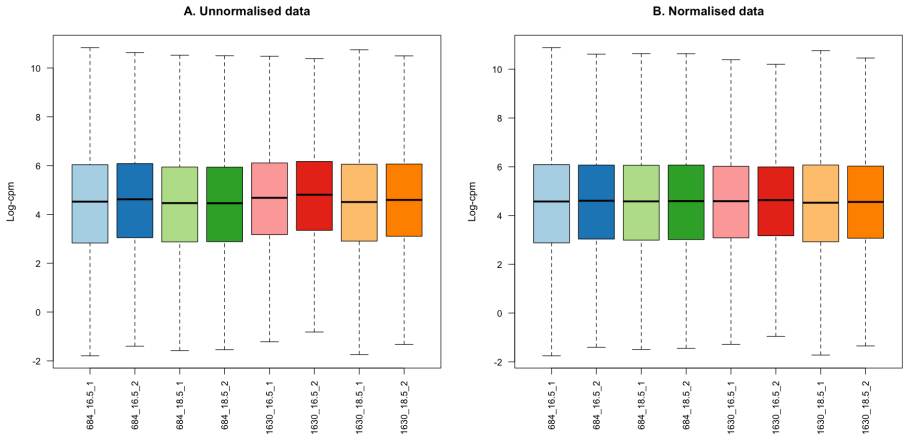

Fig. 3. Unnormalized vs Normalized 1CPM data. Left-hand side show data before applying Trimmed Mean of M-values (TMM) normalization method while right-hand side shows the same data after transformation using TMM method.

For differential expression analysis we use a linear model that assumes constant variance. Our data is discrete and violates this assumption since counts have unequal variabilities even after log-transformation. For example, on the left-hand side of Figure 4 there is a decreasing mean-variance trend for larger counts. To accommodate mean-variance relationship we estimate it from the data using robust method proposed by Law et al. [5]. To adjust for heteroscedasticity the method estimates the mean-variance trend non-parametrically and uses this relationship to predict the variance of each observation. Next it uses predicted variance to assign a weight to each observation. The weights are then used in

the linear modeling process. The right-hand side of Figure 4 illustrates the effect of adjusting the relationship of the means and the variances.

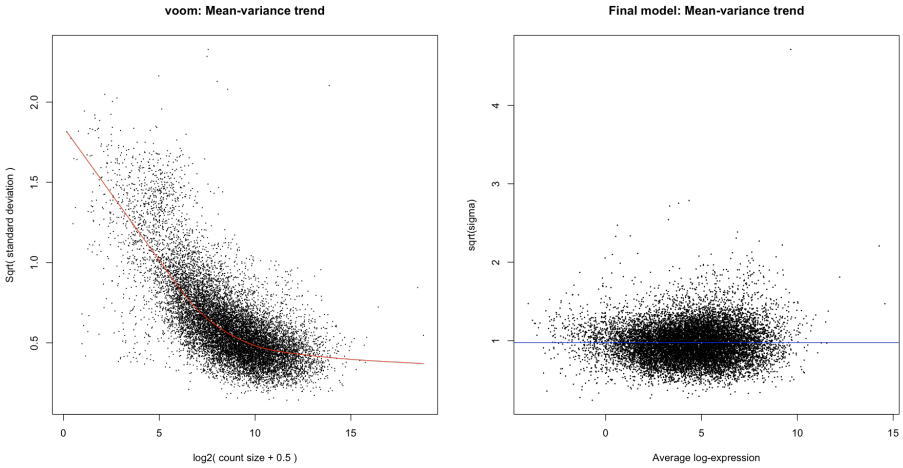

Fig. 4. Mean-variance trend before and after transformation. Each black dot represents a gene. On the left side gene-wise square-root residual variances extracted from fitting linear models to data are plotted against 1CPM counts. Red line follows estimated mean-variance trend. Right-hand side displays log2 residual standard deviations against mean 1CPM values.

Table 2. Design matrix. The matrix is set up to compare gestation ages (E16.5 vs E18.5), origin (CCHMC vs UAB) and interaction of origin and gestation age.

E16.5	E18.5	UAB	E18.5 x UAB
1	0	0	0
1	0	0	0
0	1	0	0
0	1	0	0
1	0	1	0
1	0	1	0
0	1	1	1
0	1	1	1

3.2 Linear Models, the Design Matrix, and the Contrast Matrix

To model the expression of each gene we use linear combination of different explanatory factors

$$E(y_g) = X\beta_g + \epsilon_g$$

where y_g is a response vector $y_{g1}, ..., y_{gn}$ for the gth gene, X is a design matrix representing the experimental design and βg is an unknown coefficient vector that parametrizes the average expression levels in each experimental condition. For each gene, the y_{gi} are assumed independent. We find an estimate for β_g using

$$\hat{\beta} = (X^T W_g X)^{-1} X^T W_g y_g$$

where W_g is the diagonal matrix with elements $w_{g1}, ..., w_{gn}$ derived from known counts. We assume

$$var(y_g) = \frac{\sigma_g^2}{w_g}$$

where σ_g is an unknown standard deviation and the w_{gi} are known weights. The sample variance for each gene is assumed to follow a scaled χ_d^2 distribution with d_g degree of freedom

$$S_g^2 | \sigma_g^2 \sim \frac{\sigma_g^2}{d_g} \chi_{d_g}^2$$

The unknown residual variances are allowed to vary across genes by assuming scaled inverse χ_d^2 prior distribution. The posterior variance is a combination of an estimate obtained from the prior distribution and the pooled variance

$$S_g^2 = \frac{d_0 S_0^2 + d_g S_g^2}{d_0 + d_g}$$

where d_0 and d_g are prior and empirical degrees of freedom, respectively. This way including a prior distribution of variances has the effect of borrowing information from all genes to aid with inference about individual genes. The T-statistic for a given coefficient $\beta_g i$ is

$$t_g = \frac{\hat{\beta_{gi}}}{S_g \sqrt{v_i}}$$

where v_i is the i diagonal element of $(X^T W_g X)^{-1}$

This model is described by Phipson et al. [9] and is implemented in limma [10] package, that was initially designed for analysis of data from experiments involving microarrays but was later extended and can be used for RNA-seq data as well.

In our analysis, we fit linear models to see which genes are expressed at different levels between different gestational ages and sample origins. Our design matrix is set up to help us uncover these differences as well as interactions between sample origin and gestational age. The table 2 summarizes our design

matrix. For pairwise comparisons we set up 3 contrasts: E16.5 vs E18.5 to compare gene expression at different gestational ages, UAB vs CCHMC to see how sample origin affects gene expression and E18.5 (origin UAB) vs E18 (origin CCHMC) to see whether there is any interaction between sample's origin and gestational age.

4 Ethical Considerations

Throughout the research and data analysis it is important to reflect on the ethical considerations for all involved. It is easy to feel removed from ethical concerns when examining data retrieved from third party repositories, collected by unseen researchers. This is exactly why it is imperative to refrain from simply exploring the dataset without questioning the background of how the data was collected. Understanding these processes and procedures helps ensure the collection was completed in an ethical fashion as well as assures the information gained from the analysis will be applicable to the general population.

Research involving medical experimentation requires meticulous attention to ethical concerns. Addressing those concerns often begins at the design stage of an experiment [3]. It is important to confirm individuals involved in the research experiments were treated with respect and dignity. When animals are utilized in medical research it is necessary to handle them as humanely as possible while also respecting the requirements of the experiment. Mice are largely utilized in medical studies for a multitude of reasons. Mice are small, inexpensive to obtain and care for, and reproduce quickly which allows for several generations to be studied at an expressed rate. They are also available in large quantities and can be obtained from commercial producers which breed them specifically for research purposes.

The objective of this study was to identify and characterize the developmental RNA transcriptome in alveolar epithelial type II cells of embryonic mouse lungs. The experiments used mice from the inbred laboratory strain C57Bl6. Using inbred mice lets researchers control for possible genetic variation which in turn allows the ability to have smaller test groups and leads to less mice being sacrificed. Mice are frequently used to model the biology of the human lung development as the developmental events, molecules and structures are closely similar. The results of this project are buttressed by a large body of biochemical data that allow straightforward comparisons and insight into human lung biology.

Pregnant females were sacrificed by CO_2 inhalation, individual fetuses obtained and weighed, and the lungs removed under sterile conditions. The use of CO_2 inhalation is one of the humane methods used to sacrifice animals used in medical experiments.

The ability to conduct this type of study with humans is difficult and rare. Although there is a possibility for a human subject of these gestational periods to be donated, the sample size would be extremely small. Therefore, generalizing

results to a population would not be possible. Using mice data is the most ethical possibility while still having the ability to generalize to human populations.

5 Results

To show similarities and dissimilarities between samples we used an unsupervised clustering method that is based on Principal Component Analysis (PCA). Figure 5 shows that samples cluster well within gestation age, our primary condition of interest. On the other hand sample origin also can be used to separate samples into two distinct groups. This result tells us that there is significant difference between CCHMC and UAB samples. Sex cannot be separated using any principal component.

Our model used for differential expression analysis identified 98 differentially expressed (DE) genes between E16.5 and E18.5. Of these, 50 genes were down-regulated and 48 genes were up-regulated. A mean-difference plot summarizing the DE analysis is shown in Figure 6. Figure 7 shows the number and distribution of DE genes for UAB vs CCHMC. We were surprised to find 644 up regulated and 617 down regulated DE genes when comparing samples coming from different labs. We found no genes affected by interaction of sample origin and gestational age. While most of the DE genes we found are due to difference in sample origin we found 82 genes that were different due to biological reasons, as can be seen in Figure 8 and table 3

Table 3. The top DE genes that were different due to biological reasons when comparing E16.5 and E18.5.

Entrezid	Symbol	Chromosome	PValue
12654	Chil1	chr1	0.00001813
12266	C3	chr17	0.00008442
17110	Lyz1	chr10	0.00014304
20390	Sftpd	chr14	0.00032068
71584	Gdpd2	chrX	0.00061442
72432	Spink5	chr18	0.00071023
17105	Lyz2	chr10	0.00073025
22361	Vnn1	chr10	0.00082452
170799	Rtkn2	chr10	0.00132963
16592	Fabp5	chr3	0.00132963

6 Discussion

A major advancement in molecular biology has been the development of RNA sequencing to parallel earlier advancements in DNA sequencing. The technology for creating libraries of RNA sequence data from an entire transcriptome

offers the opportunity to develop new understandings of biological processes at a more complete scale. This technological advancement has required simultaneous advancements in the capability to analyze and interpret such data. The data science capabilities for data analysis have kept apace with the developing RNA sequencing field as larger and more complex data sets are generated. In this study we have leveraged data analytical capabilities to create new insight into the development of the lung in preparation for survival after birth.

When a child is born many physiologic systems must undertake functions that they have not yet needed to perform. Healthy survival at birth is dependent on the immediate and successful implementation of some of those functions, for

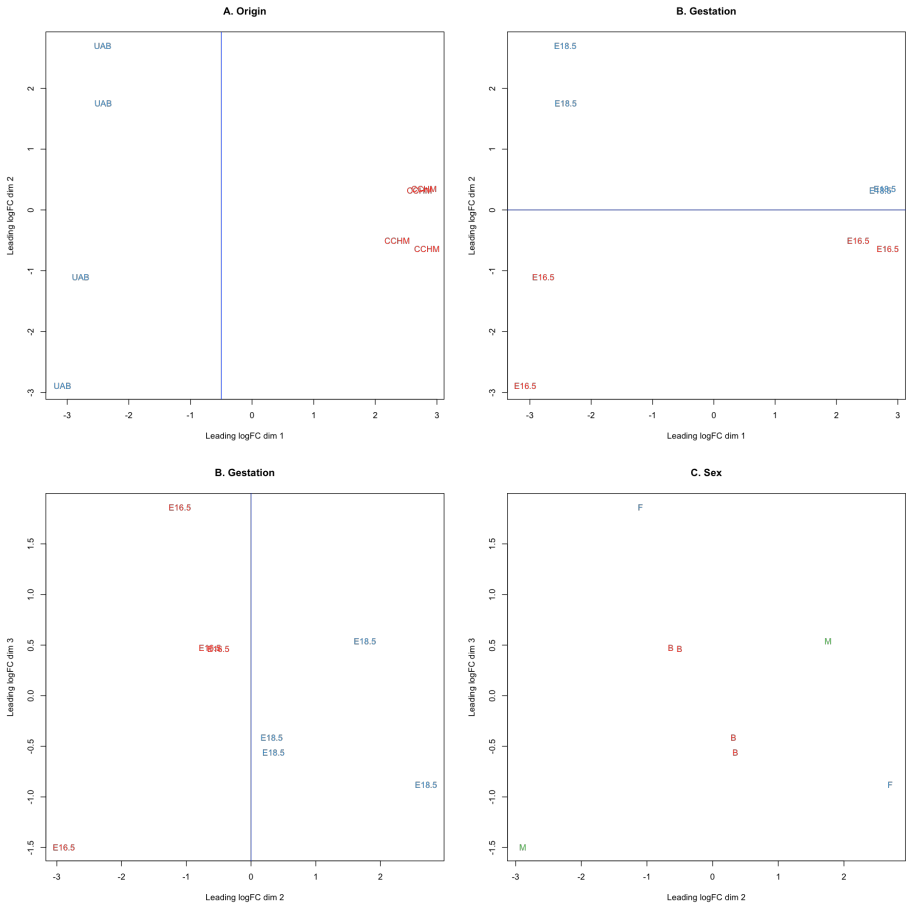

Fig. 5. Similarities and dissimilarities between samples using PCA. On the top-left there is a visible separation between sample origin using first principal component. Top-right plot shows that gestation age can be separated using second principal component. Bottom-right plot shows that there is no visible separation between different sex values.

others the successful implementation can be brought online over a period of days to months. One of the necessary immediate implementations is the ability of the lungs to provide oxygen and remove carbon dioxide via gas exchange mechanisms in the saccular and alveolar structures. These structures are lined by two types

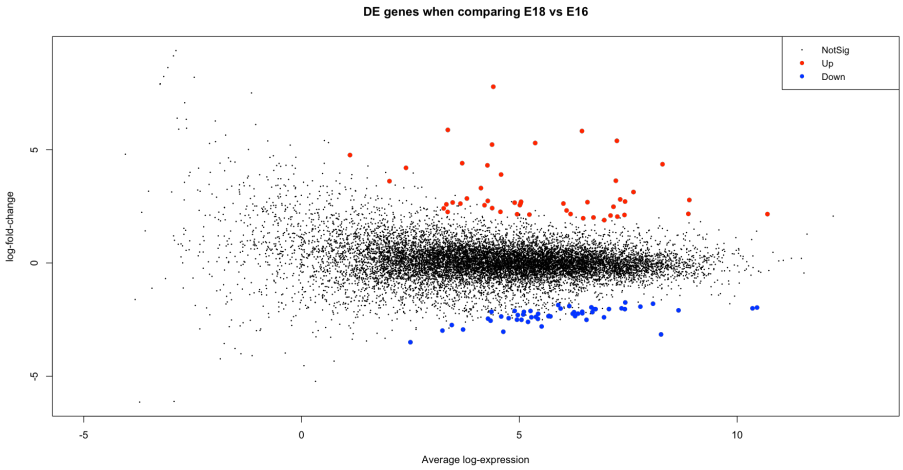

Fig. 6. Mean-difference plot summarizing DE analysis of E18 vs E16 groups which display log-FCs from the linear model fit against the average 1CPM values. Each gene is represented by a black dot. Red dots are up-regulated while down-regulated genes are shown as blue dots.

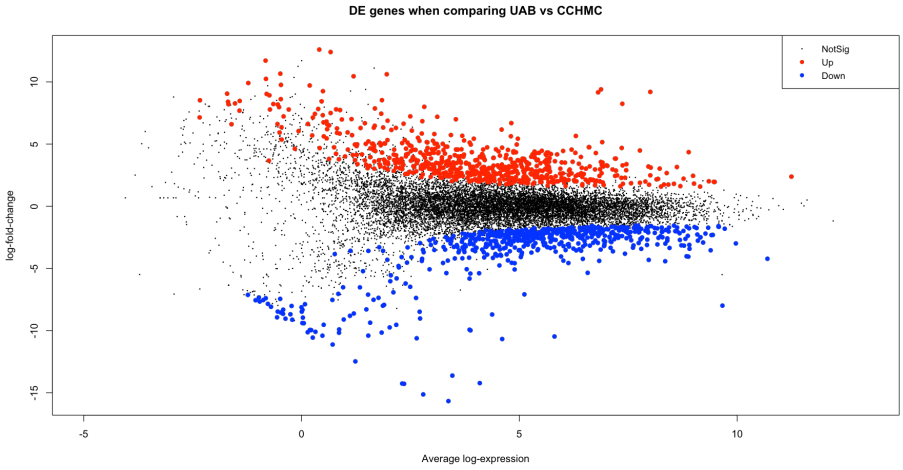

Fig. 7. Mean-difference plot displaying DE analysis for UAB vs CCHMC. Black dots are genes, red dots are up-regulated while blue dots are down-regulated genes.

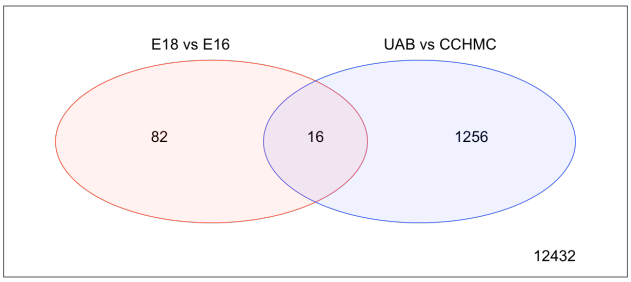

Fig. 8. Venn diagram summarizing E18 vs E16 and UAB vs CCHMC DE gene counts. We've found 12432 DE genes, 1256 were found when comparing UAB vs CCHMC, 82 when comparing E18 vs E16 and 16 genes were differentially expressed in both comparisons.

of epithelial cells, flat thin cells with large surface area, called type I epithelial cells, and taller cuboidal cells which are the AT2 cells. Each has a specific and necessary function to perform. Type I cells are the actual site of gas exchange, where gas molecules cross back and forth between the air in the alveolus and the blood stream. The physics of gas diffusion requires that type I cells have a surface that can be bathed by the surrounding gas and not obstructed by overlying layers. When a person exhales the saccules and alveoli must remain partially distended; collapse would bring opposing walls together and remove the possibility to bathe the type I cell surfaces with gas. There are physical forces that must be overcome to allow alveoli to remain distended even at the low pressures that exist with expiration. These forces include surface tension generated by the small amounts of water present in the alveolus. The surface tension forces are sufficiently strong that if not sufficiently overcome, progressive alveolar collapse will ensue and asphyxiate the person. The surface tension forces are eliminated by pulmonary surfactant, which is manufactured, packaged, and secreted by the AT2 cells.

AT2 cells mature into surfactant-producing cells after approximately 85% of gestation is completed. The maturation process includes accelerating the production of surfactant components several fold, development of the machinery needed to package surfactant, and major increases in the capacity for surfactant secretion and re-uptake. The components of surfactant, and the biology of their production, has been worked out in detail. However, there is much less understanding of other cellular processes controlling AT2 maturation.

The contribution of this study for understanding development is the identification and definition of gene expression changes that occur during the process of AT2 maturation. Changing gene expression during this developmental period has been described in many different studies, particularly for individual genes or small sets of related genes [1], but a comprehensive description of the overall numbers of changing genes has not been described. This information is needed in order to develop insight into what types of events occur in AT2 cells as they mature for surfactant production.

7 Future Work

It is important that the findings of this study be strengthened and extended by future work. The relevance of the gene profile identified here to AT2 cell maturational pathways should be pursued beginning with pathway analysis, to leverage this information into a better understanding of mechanisms of development. For example, several studies indicate gestational timing differences for AT2 maturation exist between males and females. Analysis of male and female AT2 transcript signatures could help understand why male premature infants are at greater risk of deficient surfactant production. Overall, future studies should extend our findings into additional models of AT2 development and function with the goal of identifying new approaches for treating and preventing RDS in premature infants.

References

1. Ardini-Poleske, M.E., Clark, R.F., Ansong, C., Carson, J.P., Corley, R.A., Deutsch, G.H., Hagood, J.S., Kaminski, N., Mariani, T.J., Potter, S.S., Pryhuber, G.S., Warburton, D., Whitsett, J.A., Palmer, S.M., Ambalavanan, N.a.: Lungmap: The molecular atlas of lung development program. American Journal of Physiology-Lung Cellular and Molecular Physiology **313**(5), L733–L740 (2017)
2. Bullard, J.H., Purdom, E., Hansen, K.D., Dudoit, S.: Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. BMC Bioinformatics **11**(1), 94 (Feb 2010). https://doi.org/10.1186/1471-2105-11-94, https://doi.org/10.1186/1471-2105-11-94
3. Fritzler, A.: Data science for social good (2015), https://dssg.uchicago.edu/2015/09/18/an-ethical-checklist-for-data-science/
4. Hoeing, K., Zscheppang, K., Mujahid, S., Murray, S., Volpe, M.V., Dammann, C.E., Nielsen, H.C.: Presenilin-1 processing of erbb4 in fetal type ii cells is necessary for control of fetal lung maturation. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research **1813**(3), 480 – 491 (2011), http://www.sciencedirect.com/science/article/pii/S0167488910003344
5. Law, C.W., Chen, Y., Shi, W., Smyth, G.K.: voom: precision weights unlock linear model analysis tools for rna-seq read counts. Genome Biology **15**(2), R29 (Feb 2014). https://doi.org/10.1186/gb-2014-15-2-r29, https://doi.org/10.1186/gb-2014-15-2-r29

6. Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome Biology **15**, 550 (2014). https://doi.org/10.1186/s13059-014-0550-8
7. McCarthy, D.J., Chen, Y., Smyth, G.K.: Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. Nucleic Acids Research **40**(10), 4288–4297 (2012). https://doi.org/10.1093/nar/gks042
8. Melonie Heron, P.: Deaths: Leading causes for 2017. National vital statistics reports **68**(6) (2019)
9. Phipson, B., Lee, S., Majewski, I.J., Alexander, W.S., Smyth, G.K.: Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. The annals of applied statistics **10**(2), 946–963 (Jun 2016). https://doi.org/10.1214/16-AOAS920, https://pubmed.ncbi.nlm.nih.gov/28367255, 28367255[pmid]
10. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K.: limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research **43**(7), e47 (2015). https://doi.org/10.1093/nar/gkv007
11. Robinson, M.D., McCarthy, D.J., Smyth, G.K.: edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**(1), 139–140 (2010). https://doi.org/10.1093/bioinformatics/btp616
12. Robinson, M.D., Oshlack, A.: A scaling normalization method for differential expression analysis of rna-seq data. Genome Biology **11**(3), R25 (2010). https://doi.org/10.1186/gb-2010-11-3-r25, https://doi.org/10.1186/gb-2010-11-3-r25
13. Treutlein, B., Brownfield, D.G., Wu, A.R., Neff, N.F., Mantalas, G.L., Espinoza, F.H., Desai, T.J., Krasnow, M.A., Quake, S.R.: Reconstructing lineage hierarchies of the distal lung epithelium using single-cell rna-seq. Nature **509**(7500), 371–375 (2014). https://doi.org/10.1038/nature13173, https://doi.org/10.1038/nature13173