7-11-2013

Genes of Interest in Bronchopulmonary Dysplasia: Studies of Genome-Wide Expression Profiling Data from Lung and Peripheral Blood Samples

Jared A. Mereness

Follow this and additional works at: http://scholarworks.rit.edu/theses

Recommended Citation

This Thesis is brought to you for free and open access by the Thesis/Dissertation Collections at RIT Scholar Works. It has been accepted for inclusion in Theses by an authorized administrator of RIT Scholar Works. For more information, please contact ritescholarworks@rit.edu.
Genes of Interest in Bronchopulmonary Dysplasia: Studies of Genome-Wide Expression Profiling Data from Lung and Peripheral Blood Samples

by

Jared A. Mereness

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Bioinformatics

School of Life Sciences
College of Science

Advisors and Committee Members
Dr. Tom Mariani*
Mr. Soumyaroop Bhattacharya*
Dr. Michael Osier†
Dr. Anne Haake†

*University of Rochester
†Rochester Institute of Technology

Rochester Institute of Technology
Rochester, NY
July 11, 2013
To: Head, Thomas H. Gosnell School of Life Sciences

The undersigned state that ____________________________ , a candidate for the Master of Science degree in Bioinformatics, has submitted his/her thesis and has satisfactorily defended it.

This completes the requirements for the Master of Science degree in Bioinformatics at Rochester Institute of Technology.

Thesis committee members:

<table>
<thead>
<tr>
<th>Name</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Blank] (Committee Chair)</td>
<td></td>
</tr>
<tr>
<td>[Blank] (Thesis Advisor)</td>
<td>7-11-13</td>
</tr>
<tr>
<td>[Blank]</td>
<td>7-11-13</td>
</tr>
</tbody>
</table>

Michael V. Osier, Ph.D.
Graduate Director of Bioinformatics

475-4362 (voice)
mvosier@rit.edu
ABSTRACT

The goal of this study was to identify BPD markers by gene expression profiling in lung and peripheral blood mononuclear cell (PBMC) gene expression data. Studies of BPD are often focused on lung tissue, but rarely venture to find markers of the disease outside of the most affected tissue. Genes identified from these analyses can be used as markers of BPD, and may also be linked with causal mechanisms of the disease. First, we tested the biological relevance of BPD biomarkers previously identified in lung tissue by assessing their importance in normal lung development. Next, we identified sets of differentially expressed genes at 5±2 days (Near-birth) and 25±3 days (Near-diagnosis) post birth within a publicly-available microarray dataset utilizing RNA from the PBMCs from infants at high risk for a diagnosis of BPD. These lists provided markers of BPD including ARG1, MPO, OLFM4, CEACAM1, CEACAM6 and CEACAM8 that may provide further insight into the physiology of infants with BPD. Of these, ARG1, CEACAM1 and CEACAM6 were validated by qPCR using RNA samples from lung tissue. The identified genes and their associations with the inflammatory response and vascular development via the Urea cycle provide a basis for further investigation of these pathways within BPD pathogenesis. Another benefit of these markers has been explored; using PBMC gene expression to predict long-term outcome in the near-birth time point. These studies showed that a simple 3-fold cross-validation technique could provide a model with nearly 73% accurate classification of infants at less than 1 week after birth. These studies provide a strong motive for further analyses of these genes of interest and the predictive ability of biomarkers of BPD in PBMCs.
Table of Contents

1.0.0 INTRODUCTION
  1.1.0 Prevalence and Mortality
  1.2.0 Susceptibility
  1.3.0 Normal Lung Development
  1.4.0 Disease Pathogenesis
    1.4.1 The Role of Inflammation
  1.5.0 Diagnosis

2.0.0 GENOME-WIDE EXPRESSION PROFILING

3.0.0 GOAL AND RATIONALE
  3.1.0 Specific Aim 1
  3.2.0 Specific Aim 2

4.0.0 METHODS AND RESULTS
  4.1.0 Specific Aim 1
  4.2.0 Specific Aim 2
  4.3.0 Data Extraction and Normalization
    4.3.1 Robust Multichip Average (RMA)
    4.3.2 Probe Logarithmic Intensity Error Estimation (PLIER)
  4.4.0 Quality Control and Filtering
  4.5.0 Gene Selection
    4.5.1 Significance Analysis of Microarrays (SAM)
    4.5.2 Near-Birth Samples
    4.5.3 Mid-Birth Samples
    4.5.4 Near-Diagnosis Samples
  4.6.0 Functional Classification and Pathway Analysis
  4.7.0 Molecular Validation
  4.8.0 Computational Validation
  4.9.0 Meta-Analysis

5.0.0 DISCUSSION
  5.1.0 Expression of BPD Markers in Normal Lung Development
  5.2.0 Genes Identified
    5.2.1 Myeloperoxidase (MPO)
    5.2.2 Olfactomedin 4 (OLF4M)
    5.2.3 Carcinoembryonic antigen-related cell adhesion molecules (CEACAM1, CEACAM5, CEACAM8)
    5.2.4 Arginase 1 (ARG1)
  5.3.0 Significantly Affected Pathways
  5.4.0 Genes as Early Predictors of BPD
  5.5.0 Caveats

6.0.0 CONCLUSION

7.0.0 ACKNOWLEDGEMENTS

8.0.0 WORKS CITED

9.0.0 APPENDIX A

Page
3
3
3
4
5
6
6
7
9
9
10
10
10
13
14
14
14
15
15
17
18
18
20
25
26
28
30
30
31
31
32
32
33
33
34
35
36
37
38
41
1.0.0 INTRODUCTION

Bronchopulmonary dysplasia (BPD) is a chronic lung disease of neonates, typically occurring in those who are born extremely prematurely and/or of very low birth weight (1). The disorder develops, in part, as a result of chronic exposure to mechanical ventilation and supplemental oxygen, and the damage caused by these treatments. Pathology and increased risk for lung diseases can persist through adolescence and adulthood (2). It is the most common chronic lung disease of infants in the United States, and is likely to increase in prevalence, as improved medical care increases the survival rate of premature infants (1). BPD is a highly complex disease with both genetic and environmental components, many of which need further study.

1.1.0 Prevalence and Mortality

In the US alone, approximately 500,000 infants are born prematurely every year and are at risk of developing BPD, with around fifty thousand at high risk and ten to fifteen thousand eventual cases (3). The improved ability to care for premature infants increases not only the total number of premature infants that are surviving, but also the number surviving at a greater degree of prematurity. Annually, BPD accounts for $2.4 billion in costs for infant care in the U.S. alone (3).

1.2.0 Susceptibility

BPD occurs most often in infants born earlier than 32 weeks into gestation or at a birth weight of less than 1500g (4). It is also known as Chronic Lung Disease of the newborn (CLD). Relative to the likelihood of survival, premature birth increases a newborn’s risk of developing BPD in concordance with the level of prematurity (Figure 1). However, there are many other factors that play at least as critical a

![Figure 1: Percentages of death and survival with and without BPD across a spectrum of gestational ages.](image-url)

role in an infant's predisposition. Susceptibility is now characterized by complex interactions between genetics and environmental factors (1). Recent studies examining twin-pairs have shown variation in the manifestation and severity of BPD under similar environmental conditions (1). These results provide strong evidence that genetics play an influential role in susceptibility.

13.0 Normal Lung Development

Normal lung development is severely affected in BPD. Normally, lung development occurs in several stages, both prenatally and postnatally. A total of 4 histological stages of development are completed prenatally and require a full 38 week gestational period to properly complete. While the final alveolar stage begins prenatally and occurs postnatally for up to 18 years. Lung formation enters its initial "embryonic" phase when the lungs bud from the lateral foregut endoderm between 26 and 35 days (5 weeks) of gestation. Following this is the pseudoglandular stage during which the bronchial tree

![Figure 2: Chronological diagram of normal lung development. Premature births that can potentially survive can occur between 22 to 36 weeks of gestation.](image)

expands, forming bronchi and bronchioles, spanning weeks 5 through 16. From weeks 16 through 26 the peripheral portion of the lung develops rapidly, initiating the formation of the regions in which gas exchange will occur. Beginning between weeks 24 and 26 and lasting through 36 weeks is the saccular stage. During this final neonatal stage, epithelial cells in the air spaces mature into type I (non-surfactant producing) and type II (surfactant-producing) pneumocytes, the cell types required for the most efficient gas-exchange (5). Surfactant is a fluid that increases lung compliance by reducing surface tension in the alveoli; it also prevents fluid accumulation and plays an important role in immunity and the regulation of the inflammatory response (6). Alveolarization or alveogenesis is the final stage in lung development, which begins prior to birth and continues through the first 2 decades of life. Alveoli develop and mature in this stage, dramatically increasing the surface area available for gas exchange. In parallel with this alveolar development, vascular complexity increases rapidly to facilitate gas-exchange (5, 7).

1.4.0 Disease Pathogenesis

At less than 32 weeks of gestation, infants are being born prior to the mid-saccular stage of lung development, during or prior to pneumocyte differentiation, airspace and vascular expansion and the onset of surfactant production. Birth at this stage interrupts the normal timeline of development, and lung development is considered to be arrested or delayed. The resultant deficiencies in lung function often require the use of mechanical ventilation and oxygen supplementation, as well as surfactant replacement therapy in order to oxygenate the blood to safe levels (1). Though these treatments are necessary for survival in many cases, they can be quite damaging to the delicate underdeveloped lung tissue. The consequent symptoms are highly pathologically complex, ranging from inflammation, fewer and larger alveoli, vascular dysmorphia, and aberrant accumulation of extracellular matrix. Based on clinical data, decreased lung function as a result of this damage, in conjunction with other symptoms like increased sensitivity of the inflammatory response can persist into adulthood (2, 8). Symptoms can be exacerbated by a number of factors including heredity and environmental influences like prenatal infection.
1.4.1 The Role of Inflammation

It is well-known that inflammation plays a key role in the development and progression of many pulmonary disorders, including BPD. A critical component of the inflammatory response in BPD is the presence of neutrophils and macrophages among the lung tissue (9). These two cell types accumulate in the lung, and it is thought that the neutrophils adhere to the vascular endothelium and account for some of the damage sustained to the tissue. Also characteristic of the early stages of inflammation is increased permeability of alveolar capillaries, a response that is highly correlated with a decrease in lung function due to an increased fluid volume in the lungs (10).

1.5.0 Diagnosis

Generally, the diagnosis of BPD is quite simple. BPD is defined as the need for supplemental oxygen at 28 days post birth, or as the need for supplemental oxygen at 36 weeks of gestational (postmenstrual) age (1, 2, 11). Variability in the human aspect of treatment can introduce some subjectivity to this
measure, as the "need" for supplemental oxygen may be defined differently among physicians. However, this measure's simplicity helps it to remain fairly objective. The classification of the severity of the disorder introduces many more variables. Several characteristics may be taken into account in order to determine the degree of severity. For example, a measure of the infant's fraction of inspired oxygen may be performed at 36 postmenstrual weeks for infants born before 32 weeks or at 56 days of life for those born at or beyond 32 weeks. This value is used to grade BPD severity on the scale of "mild", "moderate" or "severe" (1, 12). While these values provide grounds for diagnosis and classification, more accurate and comprehensive methods are needed.

2.0.0 GENOME-WIDE EXPRESSION PROFILING

Genome-wide expression profiling using microarrays has played a significant role in characterizing the genetic aspects of many diseases. Microarrays allow for the simultaneous evaluation of the expression of tens of thousands of genes on a single array by interrogating the composition of mRNAs expressed in a cell or tissue of interest. There have been two main platforms in this technology: glass or composite slides that are "spotted" or printed with thousands of different probe sequences or oligonucleotide probe sequences photolithographically attached to a quartz chip. The latter is pioneered by the company Affymetrix in the production of their arrays. Affymetrix GeneChips arrays use sets of oligomers as probes and each set of these probes, called a probe set, represent a gene or transcript. Expression measurements for sets of probes from individual probe sets are summarized, giving an estimate of the expression of the gene represented by the probe set. Over the years Affymetrix arrays have evolved through multiple versions. Examples of these arrays for exploring human transcriptome are the Human Genome U133plus 2.0 comprising of 54,675 probe sets representing over 27,000 unique genes and the HuGene 1.0 ST transcriptome arrays analyzing over 30,000 mRNA transcripts as well as over 11,000 long intergenic non-coding human transcripts. One of the widely used mouse arrays is the MOE430plus 2.0 array containing 45,101 probe sets representing over 19,000 mouse genes.

The experimental process for completing a microarray experiment begins with extraction of RNA from the tissue or cells, which is subsequently reverse transcribed into complementary single stranded DNA (cDNA). The underlying principle of microarray technology is 'complementary hybridization' which in this case involves binding of fluorescently labeled single stranded cDNA (or cRNA) from the tissue or cell of interest (called 'target') to the probe sequence on the chip. Biotin labeled cDNA or cRNA is then added
to the array and binds to the corresponding probe sequences. The chip is washed free of unbound material, and fluorescent stains that target biotin are added in order to visualize the results. In the case of transcriptomics, fluorescence intensities with respect to each probe set are representative of a particular mRNA's concentration or abundance, and can therefore be related to its expression level in the tissue of interest. (Figure 4)

![Figure 4: Steps in a Microarray Experiment.](http://en.wikipedia.org/wiki/DNA_microarray)

Genome-wide transcriptome profiling, enabled by microarray technology, is an extremely powerful tool in biomarker discovery. This has been effectively applied in the search for biomarkers of lung diseases. Most of the studies so far have studied expression changes in lung tissues which are typically obtained through invasive surgical procedures such as lobectomy or biopsy, which are both expensive, clinically challenging and time-consuming. In the case of BPD, lung tissue from infants can only be sampled in the event of a death. Therefore, lung tissue samples are not readily available for study, nor are they easy to obtain for testing and diagnostic purposes. In the absence of suitable human tissues, animal models have been used in order to study lung development in both the normal and diseased states (10). However, animal models cannot solve every problem, and once markers of the BPD phenotype are established in mice, lung tissue still cannot be used in human diagnosis. Therefore, other markers must be established. In contrast to lung tissues, blood samples are comparatively easier to obtain, can be drawn over several time points, and are easily processed and run. Previous studies have used whole blood-derived peripheral blood mononuclear cells (PBMC) as a means of mining for novel protein markers for lung diseases such as COPD (13, 14). PBMC are relatively easy to obtain from whole blood and contain leukocytes, including B cells, T cells, monocytes, neutrophils and natural killer cells. Application of microarrays on PBMCs can potentially provide novel biomarkers for diagnosis and therapeutic management of BPD.
Over the years, there has been a significant increase in the amount of microarray data deposited in public databases. There are several repositories for microarray expression profiling data, NCBI's Gene Expression Omnibus (GEO), ArrayExpress, and Stanford Microarray Database. These tools represent open-source databases where these data are uploaded and made available for further analyses.

Raw data from these gene expression studies are available in the form of .CEL files, which in these analyses were read into the Affymetrix Expression Console using the appropriate .CDF files (Affymetrix Inc. 2013). Each .CEL file is the output after scanning of the array and stores the expression values as intensity calculations for each sample. A .CDF file has the unique location information for each array platform.

3.0 GOAL AND RATIONALE

Knowledge of BPD and its causal mechanisms is quite limited. Through the use of gene expression profiling, this study aims to identify genes differentially expressed in both lung tissue and peripheral blood mononuclear cells, which may represent causal mechanisms of BPD. Additionally, these genes could be helpful in diagnosis of BPD. Our studies will increase the understanding of the genes and pathways associated with normal lung development that are affected by the disease.

3.1.0 Specific Aim 1

The goal of this aim was to assess genes associated with BPD for differential expression throughout lung development by examining the expression patterns of potential BPD markers in normal human and mouse lung development in lung tissue. Beginning with a list of BPD genes identified within lung tissue, we searched for differential expression of these genes during lung development. A similar study was performed using mouse models to identify genes found within the same list of BPD genes that are also differentially expressed in mouse lung development.
3.2.0 Specific Aim 2

The goal of this aim was to identify gene expression biomarkers of BPD in peripheral blood. These studies focused on a 300 sample data set obtained from the GEO, describing gene expression in peripheral blood mononuclear cells obtained from premature infants with or without a diagnosis of BPD. Blood samples were obtained from 111 subjects that are at risk for BPD at three distinct time-points. Of those, 68 (61%) were eventually diagnosed with BPD. After data preparation, expression values were examined in order to identify genes associated with BPD at each of the three time points, as well as across all three. Select markers were validated by multiple methods.

4.0.0 METHODS AND RESULTS

4.1.0 Specific Aim 1

Though there are relatively few datasets available for this investigation of BPD, those that are available have the potential to be quite informative (Table 1). Among the available datasets, many have been generated by the Mariani lab at the University of Rochester Medical Center. Two of these datasets were generated with human lung tissues using Affymetrix Human Genome U133plus 2.0 microarrays. Kho et al. (2010) identified genes as significant during human lung development by gene expression analysis (15). This data set is from normal human lung tissue samples that cover 38 unique time points from pseudoglandular and early canalicular stages of lung development (15). The Bhattacharya 2012 data set, consisting of 11 BPD lung tissues, 9 non-lung disease tissues and 8 tissues from non-BPD lung diseases, is the other(16).

In addition, the Mariani lab has generated a number of murine lung tissue studies, either describing normal lung development (15, 17) or newborn chronic lung disease (10). These data can be used to determine genes of interest within lung tissue through analyses and comparison with the human datasets. The normal mouse lung development dataset (Bonner 2003) consists of 16 embryonic and 12 post-birth samples, with 4 samples at 1, 2 and 4 weeks post birth. The datasets produced by Srisuma (2010) contain three samples from each mouse genotype (FGF KO or WT) at 1 day and 1 week post birth, as well as 3 FGF KO samples and 2 WT samples at 1 month post birth.
<table>
<thead>
<tr>
<th>Study</th>
<th>Species</th>
<th>Tissue</th>
<th>Platform</th>
<th>Sample Size</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPD</td>
<td>Human</td>
<td>Lung</td>
<td>Affy HG-U133 Plus 2.0</td>
<td>28</td>
<td>Bhattacharya 2012 AJRCM</td>
</tr>
<tr>
<td>Normal Lung Development</td>
<td>Human</td>
<td>Lung</td>
<td>Affy HG-U133 Plus 2.3</td>
<td>36</td>
<td>Kho 2010</td>
</tr>
<tr>
<td>Normal Lung Development</td>
<td>Mouse</td>
<td>Lung</td>
<td>Affy MG-U74V2</td>
<td>28</td>
<td>Bonner 2003</td>
</tr>
<tr>
<td>FGF K.O. Lung Development</td>
<td>Mouse</td>
<td>Lung</td>
<td>Affy MOE 430 Plus 2.0</td>
<td>9</td>
<td>Srinuma 2016 AJRCM</td>
</tr>
<tr>
<td>FGF WT. Lung Development</td>
<td>Mouse</td>
<td>Lung</td>
<td>Affy MOE 430 Plus 2.0</td>
<td>14</td>
<td>Srinuma 2010 AJRCM</td>
</tr>
</tbody>
</table>

Table 1: Summary of Datasets studied for Specific Aim 1.

Differential expression analysis was performed on lung tissue samples of the Bhattacharya 2012 dataset in order to generate BPD markers in human lung tissue. This study found a set of 159 genes that were differentially expressed between the two classes (16). The Kho 2010 study found 3224 genes that were differentially expressed during normal human lung development by principal component analysis (PCA).

Genes in the Bonner et. al. (2003) dataset was analyzed for significant differential expression during postnatal development by Analysis of Variance (ANOVA). ANOVA generates a test statistic in the form of a p-value for each gene based on the assumption that all means are random samples of the same data. A p-value equal to or less than 0.05 means that the variances are not equal and therefore, that the expression of the gene differs significantly at some point during postnatal lung development (18).

Gene expression results from FGF knockout and WT mice in the Srinuma et. al (2010) study were compared via t-test at three past birth time points. These time points were 1 day, 1 week and 1 month post-birth. The t-test generates a t-score for each gene based on the difference in the means of the diseased and control groups. Any given gene will be considered significantly different between the groups by t-test if the calculated p-value is less than or equal to 0.05 (19). This will be performed in Microsoft Excel 2010.

The 159 genes identified by Bhattacharya et. al. were compared with genes of interest identified in all other datasets. Comparisons of the datasets were performed based on gene symbol in the Database platform of Apache OpenOffice 3.

Of the 159 genes identified in lung tissue of Human BPD, 32 were found significant by ANOVA in the Bonner et. al. (2003) dataset examining normal mouse lung development (Figure 3). In the Kho et. al. (2010) dataset, 57 of these 159 BPD genes were identified as significantly different during human lung
development (Figure 6). In the Srivastava et al. dataset, none of the significant genes identified by t-test at any of the three post-birth time points were present in the human BPD gene list. Between the 32 BPD genes identified in mouse lung development and the 57 identified in human, 21 were common (Table 2).

![Diagram 1](image1)

**Figure 5**: Genes identified as significant by ANOVA that match between 159 gene list from human BPD. Performed on postnatal data.

![Diagram 2](image2)

**Figure 6**: Genes matching between 159 genes known to be differentially expressed in human lung tissue and those observed to be significant by PCA throughout normal lung development.

<table>
<thead>
<tr>
<th>Gene</th>
<th>stats</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOP2A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMD2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RBP4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>POSTN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NEK2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MKI67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPS23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IGAP3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECH2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IGF1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>COL4A1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CENPA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDKN3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDC26C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BUB1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIRC5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASPM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>APOH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANLN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGTRB</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 2**: List of genes overlapping between the 32 identified as significant in mouse lung development and BPD lung tissue, and the 57 identified as significant in human lung development and BPD lung tissue.
4.2.0 Specific Aim 2

These studies are based on a human peripheral blood microarray data set currently available in the Gene Expression Omnibus (GEO). This dataset includes expression values from 299 peripheral blood mononuclear cell RNA samples collected from around 300 human subjects at three time-points, 5±2 days (near-birth), 15±2 days (midpoint), and 25±3 days (near-diagnosis) days after birth, run on HuGeneST 1.0 arrays (22). Samples were obtained at the Oslo University Hospital in Norway and are well-documented with numerous phenodata accompanying each sample.

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Near-Birth 5±2 days</th>
<th>Midpoint 15±2 days</th>
<th>Near-Diagnosis 25±3 days</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (Control)</td>
<td>35</td>
<td>40</td>
<td>37</td>
<td>112</td>
</tr>
<tr>
<td>1 (Mild)</td>
<td>34</td>
<td>34</td>
<td>38</td>
<td>106</td>
</tr>
<tr>
<td>2 (Moderate)</td>
<td>13</td>
<td>11</td>
<td>10</td>
<td>34</td>
</tr>
<tr>
<td>3 (Severe)</td>
<td>15</td>
<td>13</td>
<td>14</td>
<td>42</td>
</tr>
<tr>
<td>NA</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 3: Distribution of samples across each of the three time points represented by the Nygard 300 sample dataset.

As outlined in Figure 7, using the available raw data files, expression intensities were extracted and normalized using multiple approaches, followed by standard quality control procedures. Samples and genes passing quality control thresholds were used for differential gene expression analysis. Differentially expressed genes identified by multiple analytical approaches were subsequently used for pathway analysis, validation, and to test the efficacy and feasibility of subject outcome prediction based on gene expression profiles.

![Figure 7: Outline of data analysis](image-url)
4.3.0 Data Extraction and Normalization

4.3.1 Robust Multichip Average (RMA)

The RMA algorithm is a widely used normalization approach and is used to generate an expression summary and metrics at the probeset level by using a linear model for the estimation of the signal and the error. The probeset or transcript level expression values are computed as background corrected, Log2 transformed and quantile normalized values (21, 22). When performed across all arrays, this makes the comparison of expression values between any set of arrays quite simple to perform. One caveat to this method is that all samples must be normalized at once (23). For the following analyses, samples were grouped by time point and each group was normalized separately.

4.3.2 Probe Logarithmic Intensity Error Estimation (PLIER)

PLIER produces probe-level expression values, and while summarizing at the probe-set level, it accounts for experimentally observed patterns of signal intensity. The error model used in PLIER makes the assumption that error is proportional to observed intensity and ignores the background intensity. This approach allows PLIER to be more sensitive to differences in expression at near background levels, and to handle errors more accurately than many competing algorithms (24). Though there are some distinct advantages and considerations included in the PLIER algorithm, there are some potential disadvantages. The PLIER algorithm tends to over-estimate gene level expression changes between arrays, especially for probe sets with high expression values (23).

Raw data were extracted and normalized using both of these algorithms, Probe Logarithmic Intensity Error Estimation (PLIER) and Robust Multichip Average (RMA). This yielded a matrix consisting of 299 samples and expression values for 33,297 probesets representing 28,869 distinct genes for each normalization technique. All normalization was performed using Expression Console version 1.2.1 (Affymetrix, Inc. Santa Clara, CA).

4.1.0 Quality Control and Filtering

Quality control is an essential component of an expression profiling study. It ensures that high quality data is used in the study and that no confounding factors are present. For this project we applied multiple quality control thresholds at the sample level on all the datasets.
The first step was visual inspection of array images for defects or abnormalities. This was performed using Affymetrix Expression Console. This was complemented by generating intensity distributions and hierarchical clusters for all samples at each time point to check for anomalies that may have been indicative of issues with a particular sample. Samples were grouped by time point to examine how the median signal intensity and width of signal distribution compared with all other samples in the group. Hierarchical clustering was performed using average linkage clustering based on a Euclidian distance measurement. Signal distribution plots were generated in Affymetrix Expression Console. Hierarchical clustering was performed in MultiExperiment Viewer (MeV) version 4.8.1.

Visual inspection of the sample images did not yield any problematic samples; however, histograms and hierarchical clusters generated using samples separated by time point, based on all genes (Appendix A) provided further information. From near-birth samples, the signal distribution from one sample was much broader than most other samples, and the sample also did not group with any other samples and was therefore eliminated from further analyses. All samples without a diagnosis were also eliminated from further analysis.

4.5.0 Gene Selection

Samples and genes that were retained after quality control procedures and filtering were then assessed for differential gene expression. The first test for differential expression was Significance Analysis of Microarrays (SAM). The second criterion to be used for differential expression was Welch’s t-test for unequal variance coupled with fold change. Both the tests were applied on both RMA and PLIER normalized datasets.

4.5.1 Significance Analysis of Microarrays (SAM)

Developed by Tusher et al. in 2001, Significance Analysis of Microarrays (SAM) is a widely used method for identifying differentially expressed genes from microarray data. One feature of SAM is that it provides an estimation of the False Discovery Rate (FDR), which is the estimated number genes likely to be false-positive. SAM also allows for a great deal of tuning (through parameter delta) to increase or decrease the threshold, and therefore sensitivity of the analysis.
For any given gene, a score (d-value) is calculated which is based on difference between the mean expression values of the groups divided by the standard deviation of the measurements. Genes are then ranked according to their d-value, which is designated as the observed d-value for each rank. Then labels of the samples are randomized between groups 1 and 2 by permutation and the d-value for each gene is calculated again for this randomized set and ranked. This permutation process is repeated a designated number of times and the average d-value for each rank over all permutations, is calculated, which is the expected d-value for each rank. The observed d-values are then plotted against the expected d-values for the corresponding ranks in a SAM plot. The delta value is the threshold amount by which a gene’s observed d-value must deviate from the expected d-value in order to be considered significant. The more a gene’s observed d-value deviates from the expected, the more likely it is that the gene is truly differentially expressed (25). SAM analysis was performed using MultiExperiment Viewer (MeV) version 4.8.1.

![Figure 9: Steps in SAM. Image courtesy of Saffra Shahal](image)

Genes were identified as significant if they were significant by SAM analysis at a median FDR of 0, or by t-test p-value less than 0.05 and a fold change of greater than 2 or less than one-half. The t-test method has been described previously in the Specific Aim 1 methods; section 4.1.0. For all genes, a fold change value was calculated by dividing the mean expression value of all the diseased samples in a group by the mean expression values of the control samples. For PLIER-normalized data, the threshold (1), was set such that genes with ratios greater than or equal to 2, and less than or equal to 1/2 were considered significant according to fold change. In the case of the log2-corrected expression values generated by RMA-normalization, affected expression values were subtracted from rather than divided by the control.
value. A threshold of 1 and -1 were used, indicating a doubling or halving of expression. T-test and fold change calculations were performed using Microsoft Excel 2010.

4.5.2 Near-birth samples

SAM analysis of PLIER normalized data yielded 963 genes that were significant when comparing control samples with all BPD samples (n=96), while the combination of t-test and fold change of 2 or greater found 26 significant genes. A total of 16 genes were significant by all criteria. There were 385 genes found to be significant by SAM when comparing control samples to only severe BPD samples (n=49), and 62 genes were significant based on t-test and fold-change, with a total of 20 genes significant by all criteria (Table 4a).

Examination of RMA-normalized data by SAM gave 1011 significant genes when comparing controls and all BPD samples (n=96), with 42 significant by t-test with a 2-fold or greater difference between groups. A total of 41 were significant in all categories. Comparison of control and severe BPD samples (n=46) provided 426 genes significant by SAM and 163 significant by t-test and fold-change, with 62 meeting significance criteria for all three tests (Table 4a).

Comparison of the genes identified as significant between RMA and PLIER when examining control and all BPD samples yielded 15 similar genes between groups. An analogous comparison when examining control and severe BPD samples only gave 18 significant genes similar between RMA and PLIER normalization (Table 4a). Between the 15 and 18 gene lists, 10 were shared (Figure 9).

<table>
<thead>
<tr>
<th>Near-Birth</th>
<th>PLIER</th>
<th>RMA</th>
<th>Matching</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control vs. All BPD</td>
<td>16</td>
<td>41</td>
<td>15</td>
</tr>
<tr>
<td>Control vs. Severe BPD</td>
<td>20</td>
<td>62</td>
<td>14</td>
</tr>
<tr>
<td>Combined</td>
<td>10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Near-Diagnosis</th>
<th>PLIER</th>
<th>RMA</th>
<th>Matching</th>
</tr>
</thead>
<tbody>
<tr>
<td>Controls vs. All BPD</td>
<td>33</td>
<td>35</td>
<td>19</td>
</tr>
<tr>
<td>Control vs. Severe BPD</td>
<td>51</td>
<td>214</td>
<td>58</td>
</tr>
<tr>
<td>Combined</td>
<td>18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiclass (0.3 Diagnosis)</td>
<td>859</td>
<td>13 Matching Combined</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Summary of significant genes identified by each normalization method and those matching between each method as well as those matching between analyses of controls and all BPD compared with analyses of controls and severe BPD samples at the near-birth time point (A) and near-diagnosis time point (B).
4.5.3 Midpoint Samples

Midpoint samples provided 11 significant genes only when comparing Controls with Severe BPD samples. This result, combined with the generally wider signal intensity histograms provided grounds for the removal of these samples from further analyses.

4.5.4 Near-Diagnosis Samples

Analysis of PLIER normalized samples found that 1115 genes were significant when comparing control samples to all BPD samples (n=99) with SAM, and 63 significant genes based on t-test and the fold-change threshold of 2. There were 33 genes significant by all of these criteria. Comparison of controls and severe BPD samples (n=51) yielded 2009 genes significant by SAM, and 121 that were significant by t-test and fold-change. A total of 51 genes were significant by all tests. Multiclass SAM analysis of these samples yielded 859 significant genes (Table 4a).

Analysis of RMA-normalized samples by SAM yielded 1220 genes significant by SAM when comparing controls to all BPD samples (n=99). There were 36 genes found to be significant by t-test and fold-change threshold and a total of 35 were significant by all three measurements. Comparisons of control and severe BPD samples (n=51) gave 1797 genes significant by SAM and 214 significant by t-test and fold-change threshold. A total of 214 were significant according to all criteria (Table 4b).
Examination of the significant samples identified in the PLIER and RMA normalized data when comparing controls and all BPD samples yielded 19 matching genes. A similar comparison with control vs. severe BPD samples provided 38 similar genes (Table 4b). Between these 19 and 38, 18 were shared within the 38 (Figure 10). Of these 19, 13 were also present in the multiclass SAM results (Figure 11). Of all genes identified at either the near-birth time point or the near diagnosis time point, there were three genes matching between the two groups (Figure 12). Between the 18 genes overlapping at the near-diagnosis time point and the 10 overlapping at the near-birth time point, one gene (ARK1) was similar.
4.6.0 Functional Classification and Pathway Analysis

Functional classification was performed in order to determine the over-represented ontological classes in groups of genes found to be significantly differentially expressed. These analyses were performed using Database for Annotation, Visualization and Integrated Discovery (DAVID). Canonical pathways significantly dysregulated in the disease system were identified using Ingenuity Pathway Analysis (IPA) which obtains its pathway information from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Ingenuity®Systems, Inc. Redwood City, CA). Comparative analyses of the expression of differentially expressed genes across multiple tissues were examined using BioGPS.

DAVID is a free online resource that allows for functional classification and Gene Ontology (GO) enrichment analyses. Functional classification using gene ontology was performed to provide information on differentially expressed genes associated with different biological processes, molecular functions and cellular components. GO classes with p-value less than 0.05 were defined as significantly overrepresented.

IPA software is commercially available large repository of canonical pathways and biological interactions. IPA was used to obtain canonical pathways significantly dysregulated in the differentially expressed genes. Pathways with p-value less than 0.05 were defined as significantly dysregulated.

BioGPS is a free gene annotation portal that allows for further functional classification of genes of interest. For many genes, there is information on relative expression across many different tissue and...
cell types. Genes of interest identified by PBMC analysis were examined with this tool to investigate their relative expression in lung tissue.

Gene ontology analysis of near-birth genes in David found 9 significant classes associated with the 15 genes identified by analysis of controls and all BPD samples, and 14 classes associated with the 16 genes identified when comparing controls and severe BPD samples only (Figure 13).

A similar analysis of near-diagnosis genes in David found 7 significant classes associated with the 19 genes identified by analysis of controls and all BPD samples, and 17 classes associated with the 38 genes identified when comparing controls and severe BPD samples only (Figure 14).

Figure 13: Results of David analysis of near-birth gene lists. The “All genes” measurement is the percentage of genes involved in the pathway identified within the gene list. The “BPD Markers” measurement is the percentage of genes in the input list that identify with the particular pathway. Classes outlined in red represent those present in both analyses.
Figure 14: Results of DAVID analysis of near-diagnosis gene lists. The “All genes” measurement is the percentage of genes involved in the pathway identified within the gene list. The “BPD Markers” measurement is the percentage of genes in the input list that identify with the particular pathway. Classes outlined in red represent those present in both analyses.
Further analysis of canonical pathways via IPA show 7 pathways linked to the 15 near-birth genes obtained by comparison of controls and all BPD samples and 27 pathways linked to the 19 near-diagnosis genes found when comparing controls and all BPD samples. The near-birth sample analysis showed several pathways linked to melatonin and arginine degradation and citrulline biosynthesis (Figure 15). In the near-diagnosis time point, similar pathways are present and are joined by many related to other inflammatory response and disease pathways, including chronic obstructive pulmonary disease airway pathology (Figure 16).

![Canonical Pathways affected in All BPD near-birth](image)

**Figure 15:** Significantly affected pathways in control vs. all BPD gene list from near-birth samples identified via IPA. Shown are the $-\log(p$-value$)$ of the identified pathways. Pathways outlined in red are also present when analyzing the controls vs. severe BPD gene list.
Figure 16: Significantly affected pathways in control vs. all BPD gene list from near-diagnosis samples identified via IPA. Shown are the -log_{p-value} of the identified pathways. Pathways outlined in red are also present when analyzing the control vs. severe BPD gene list.
4.7.0 Molecular Validation

The quantitative, real-time, reverse-transcriptase-polymerase chain reaction method (qPCR) was used for molecular validation of differentially expressed genes. This technique was performed in order to quantify the abundance of a specific RNA molecule. Similar to microarrays, the first step involves reverse-transcriptase (RT)-based copying all RNAs present in a sample into cDNA. The second step involves Polymerase Chain Reaction (PCR)-based amplification of a specific cDNA to define the abundance of the mRNA it represents. The amplification was measured in real-time, allowing the continuous quantification of product. The approach used for product detection involved the use of fluorescent dyes that bind to all double-stranded DNA in the sample, whereby an increase in total DNA content of the sample causes a quantifiable increase in fluorescence (26). The qPCR validation of genes identified in PBMC data was performed using lung tissue samples from control and BPD subjects.

There were seven candidate genes chosen from the results of the human PBMC sample analyses for molecular validation in human lung tissue. The qPCR analysis was performed using 11 BPD lung tissue samples and 8 control lung tissue samples; these were the same control and BPD tissue samples used by Bhattacharya et al. (2012) to determine the list of 159 lung tissue BPD markers by expression profiling. One of the 9 control samples used in the Bhattacharya et al. study was not used due to an insufficient volume of cDNA. Of the seven genes examined, three were successfully validated as significantly differentially expressed by either t-test or Mann-Whitney U test. AEG1 expression in BPD lung tissue was significantly increased via t-test ($P=0.033$) and also showed a trend toward significance by Mann-Whitney U test ($P=0.073$). Expression of CEACAM1 was observed to be significantly increased in BPD by Mann-Whitney U test ($P=0.027$) with a trend toward significance via t-test ($P=0.052$). The expression of CEACAM6 was significant by both t-test ($P=0.042$) and Mann-Whitney U test ($P=0.001$). The expression of MPO showed a trend toward significance by t-test (0.064) (Figure 17).
4.2.3 Computational Validation

Computational validation was performed using the 3-fold cross-validation approach on data from the near-birth time point. For this analysis, the dataset of a particular time point was randomly split into 3 groups; two groups were combined and used as a training set. A binary logistic regression model was applied based on genes identified as significant within the training set by SAM, and the model was tested against the removed testing set. Binary logistic regression modeling is used to predict a binary (2-class) outcome. For this analysis, samples were grouped into two categories: controls and BPD (arrest diagnosis). The logistic regression modeling was performed in Minitab 16 (Minitab, Inc. State College, PA). This was repeated 3 times, once for each combination of groups, and cumulative percentage of correct predictions based on each of the three models provided a measure of the efficacy of the model. The genes identified as the best predictors in these models were compared with the genes of interest identified from expression profiling for validation (27).

Three-fold cross-validation via binary logistic regression modeling performed on PUEH normalized data from near-birth samples yielded similar results across the range of test set sizes. A model using the 8
most significant genes identified after partitioning of the samples provided the best predictive ability of Control or BPD sample with 72.63% accuracy (Figure 18). However, models between 3 and 10 predictors showed differences in prediction of ±1 correctly classified sample. This accuracy slowly decreased with the addition of more predictors with the lowest predictive ability of 60.00% coming from the largest group of predictors, containing 32 genes. Of the 20 top predictors identified by any model (a total of 44 genes), 16 were present in the list of 58 genes. Included in this list were ARG1, MPO, OLFM4, CEACAM6 and CEACAM8 (Figure 19).

Figure 18: Percentage of correct simple classifications based on the number of predictors used in each model using 3-fold cross-validation.

Figure 19: Genes identified in the top 10 predictors for any of the 3 models that were also present in the list of 58 significant genes identified in either comparison at either time point. The highlighted gene was also a candidate used in one of the 3-gene models.
4.9.0 Meta-Analysis

The list of 58 genes that was identified in the gene selection section based upon significance either near or below a diagnosis via control vs. all BPD or control vs. severe BPD analyses was then subjected to a similar set of comparisons as the 159 gene list from the Bhattacharya 2012 study. Methods for this analysis have been previously described under Specific Aim 1.

Of the 58 genes identified in the PBMC analyses, 29 were found in the mouse lung development dataset, of which 16 were found to be significant by ANOVA of postnatal development data points (Figure 20). Further, 15 of these 58 genes were identified as significantly different during human lung development (Figure 21). Of the 42 genes identified as significant at any postnatal time point in the FGF KO vs. WT datasets, 34 of them were also present in the 58 gene list from the PBMC analyses (Figure 22). Of these 58 genes, 2 were also found in the list of 159 potential BPD markers identified in lung tissue (Figure 23). These two genes were CEACAM6 and CD177.

![Diagram](Image)

**Figure 20:** Genes identified as significant by ANOVA that match between 58 gene list from PBMC performed on postnatal data.
Figure 21: Genes matching between 58 genes found to be differentially expressed in PBMCs and those observed to be significant throughout normal lung development.

Figure 22: 42 genes from the 58 gene list were found in the FGF/KO dataset. Of the 42, 34 were found to be significantly different between FGF and WT by T-test at either 1 day (26 genes), 1 week (26 genes), or 1 month (17 genes) post birth.
5.0.0 DISCUSSION

5.1.0 Expression of BPD Markers in Normal Lung Development

Meta-analysis performed on lung development datasets in section 4.1 showed that there are genes that are affected by BPD and also influential during the development of lung tissue. As BPD is a developmental disorder, a study examining genes that are related to lung development that are also affected in BPD could provide the best markers and the most insight into the causal mechanisms of the disease. The gene lists identified can also help interpret the results of the PBMC data by relating what is occurring in the lung tissue to what is observed in PBMCs. These smaller gene lists may help narrow some of the results obtained by these studies and focus on what may be more promising genes within these larger lists.

The genes identified as matches between these datasets are potential markers of the disease within lung tissue, and may contain links to causal mechanisms of the disease. These data provide insight into the mechanisms affected directly within the lung tissue. Detailed annotation of the genes identified by these comparisons show an overwhelming number of classes that are related to cellular division and growth, and may affect. Similar to the genes identified, these are classes that are involved in lung 

Figure 23: 2 of the genes identified in PBMC data were also identified in BPD lung tissue in the Bhattacharya 2012 study. The two genes overlapping were CEACAM6 and CD177.
development that are also affected in BPD (data not shown). This result shows that the goal of assessing BPD markers for involvement in lung development has been met, and that further analysis of the roles of these genes within normal development as well as in the diseased state may provide much information on the pathogenesis of BPD.

5.2.0 Peripheral Blood Markers of BPD

The observation that more genes were significantly different in near-diagnosis (25±3 days post-birth) than in near-birth (5±2 days post-birth) is interesting and agrees with much of what is known about the disease. Near birth, little or no physical distinction can be made between these premature, low birth-weight individuals, only that they are all “at risk” of developing BPD. It follows that there may be less variation in gene expression between affected individuals and non-affected individuals at this point, as the disease has not progressed and the environmental components have not been allowed to develop. However, genetic differences at this early time point may be more likely to be linked with causal mechanisms of the disease, whereas additional differences at the later time points may be driven by these mechanisms or environmental conditions, and are more likely to be markers of the disease.

Of all BPD genes identified from near-birth and near diagnosis, regardless of the method of comparison, there were three genes in common: Arginase 1 (ARG1), Myeloperoxidase (MPO), and Olfactomedin 4 (OLFM4). Of all four comparisons, control vs. all BPD and control vs. severe BPD at both time points, ARG1 was identified as significant in all four, MPO was found to be significant in three of four and OLFM4 was significant in two of the four (Figure 12). There were a total of 58 genes identified as significant at either time point via either method of comparison; this list was considered to be the list of BPD gene markers in PBMCs identified in this study. These 58 genes also contained CEACAM1, CEACAM6 and CEACAM8. CEACAM6 was determined by Bhattacharya et. al. (2010) to be differentially expressed in BPD lung tissue, providing a strong motive for further examination of these three genes.

5.2.1 Myeloperoxidase (MPO)

The results of gene expression comparisons among tissues by BioGPS show that MPO is highly expressed in CD34-positive cells; CD34 is highly expressed in fetal lung tissue (29). Many studies have shown that MPO is a marker of neutrophil activity and presence. It has been shown that a measure of MPO activity is directly positively correlated with neutrophil presence in the immediate vicinity in a particular tissue (30). Neutrophil accumulation is well characterized as a part of the inflammatory response, a major component of the pathology of BPD. Interestingly, it has been found that neutrophil accumulation in
the lungs of children with BPD following the onset of mechanical ventilation results in decreased numbers of circulating neutrophils within minutes (9). This finding, coupled with an increased presence of MPO in PBMCs as observed in this gene expression study may suggest that MPO is not only a marker of neutrophil presence, but that it’s other roles may also warrant further investigation.

5.2.2 Olfactomedin 4 (OLFM4)

There are several studies that have shown that OLFM4 is an important factor in controlling cellular growth and apoptosis. It has been observed that OLFM4 has growth-promoting and anti-apoptotic properties in vitro (31). The observed increase in OLFM4 expression may be a response to the lack of growth on a systemic level due to prematurity of the infant. Studies have also shown that OLFM4 has a direct role in mediating neutrophil activity. Decreased OLFM4 expression leads to an increase in neutrophil-mediated killing during an infection in mice (32). An increase in OLFM4 as observed in the PBMC expression data may be a response to increased neutrophil activity in the blood or somewhere else in the body. The observed increase in OLFM4 may be related to increased neutrophil presence or activity. Similarly to MPO, OLFM4 may also be a marker for the presence of neutrophils.

5.2.3 Carcinoembryonic antigen-related cell adhesion molecules (CEACAM1, CEACAM6, CEACAM8)

Similar to the role that OLFM4 plays, the CEACAM family of genes influences cellular differentiation, structural arrangement and apoptosis (33). They also play a role in the immune response, as CEACAM1, CEACAM 6 and CEACAM 8 among others have been implicated in Neutrophil activation and adhesion (34). Though these three genes were identified as significant through the analysis of PBMC data, they were also validated as significantly different between BPD and non-BPD control samples of lung tissue via qPCR. CEACAM6 was also observed to be differentially expressed in BPD lung tissue by Bhattacharya et. al (2010). This is especially interesting as data obtained from the BioGPS database shows a high relative expression of CEACAM6 in lung and tracheal tissues (35). As CEACAM6 was identified as differentially expressed in PBMCs at the near-birth time point, CEACAM6 or the CEACAM family of genes may be involved in the development of BPD. These genes could also play a similar role to OLFM4, and may be upregulated for the same reasons: the lack of development in certain tissues due to premature birth or to promote inflammation.
5.2.4 Arginase I (ARG1)

The presence of ARG1 as a differentially expressed gene in all four major gene selection analyses spanning both the early and late time points prompted further investigation. The expression of ARG1 within macrophages has long been thought to be a promoter of inflammation, fibrosis and wound healing. All of these are major issues happening within the lungs of infants with BPD. However, a recent study in mice found that ARG1 functioned as an inhibitor of fibrosis and inflammation following infection, and that mice lacking ARG1 expression died sooner of infection (36). Important to note is the fact that human macrophages do not express ARG1 (37). The role of Arginase I could be similar in humans, however, the source must be different. Interestingly, according to expression data from BioGPS, ARG1 is significantly expressed in very few tissues, fetal lung tissue being one of the few (38). This observation suggests that as ARG1 is highly expressed in fetal lung and is also differentially expressed in BPD from the near-birth time point through the near-diagnosis time point, it may be involved in a causal mechanism of the disease.

An alternate role that ARG1 may play, especially in the development of fetal lung tissue is in the process of vascularization. A new study focusing on retinopathy, a common condition of premature infants involving the vascular tissues of the retina, may be related to increased ARG1 expression (39). An increase in the expression of ARG1 decreases the availability of L-arginine to nitric oxide synthase which causes a decrease in nitric oxide and increases oxidative stress. Arginase I also causes the formation of polyamines and proline, which induce cell growth and inflammation of vasculature. This study describes the potential role that ARG1 may have in the vascular dysfunction in retinopathy and the potential for treatment of the condition by inhibition of ARG1 activity (39). Arginase I may play a similar role in the vascular dysmorphia and increased sensitivity of the inflammatory response observed in BPD. The results of qPCR in lung tissue provide further evidence that it may be a key part of the mechanism driving this condition. Moreover, iNOS scavenges nitric oxide, which could exacerbate the problems caused by already low levels of L-arginine.

5.3.0 Significantly Affected Pathways

From the analysis of the particular genes of interest in section 5.2, it seems as though genes linked with the inflammatory response are also quite closely associated with cellular growth, differentiation, and vascularization in particular. It is clear that genes and pathways involved in the inflammatory response are significantly affected in the PBMCs of individuals with BPD, with increasing involvement as the
infants develop through the first four weeks of life. Interestingly, the three genes that were consistent between both time points have previously been identified as markers of specific inflammatory cells, including neutrophils and macrophages. MPO has been identified as a marker of neutrophils, and OLSM4 has also been linked with their presence (40). Analysis of the significant gene list from near-birth samples using David shows many classes related to the inflammatory response and immune cell activity, with a few classes linked to components of the cell membrane (Figure 13). Similar analyses of near-diagnosis data show a shift in this balance, with a large number of classes associated with membrane components, and a few others corresponding to the inflammatory response (Figure 14). Genes like those in the CES6CAM family make up several of the genes identified as membrane components, and their other roles have been identified in section 5.2.3. Finally, there were also classes identified in the gene lists entered at both time points that involved carbon-nitrogen hydrolase activity, which likely point to the activity of ARG1 and other components of the urea cycle. The gene ontology information obtained is highly consistent with the activity of the genes of interest identified above.

Canonical pathway analysis also provided highly consistent data with the identified functions of the major genes of interest. In the near-birth samples, there is an overwhelming presence of pathways related to arginase degradation and the urea cycle. This is highly suggestive of significant involvement of ARG1 and related genes in BPD cases. These pathways linked with the activity of arginase are again present in the genes identified in the near-diagnosis samples, though they are somewhat diluted within the list of 27 significant pathways. This is not unexpected, as the disease has progressed and begun to affect more systems at the near-diagnosis time point. These, along with pathways involving inflammatory cells provide consistency within the pathways identified from PBMC gene lists and what is known about the issues surrounding the lung tissue in BPD, such as increased sensitivity of the inflammatory response.

5.4.0 Genes as Early Predictors of BPD

Currently, the diagnosis of BPD is a somewhat subjective measure that is not made until many weeks after birth. The ability to predict the outcome or diagnosis of a premature newborn within days of birth would be a very powerful tool. A successful predictor based on the expression of a few genes that could be obtained from a simple blood test rather than a tissue biopsy would make this a very feasible and desirable method. This would allow certain actions to be taken sooner and may open up new treatment options and opportunities. The cross-validation modeling was performed at the near-birth time point for exactly that reason. The baseline success rate for a model containing zero predictors for this analysis.
was 64%, which is the accuracy that would be obtained by classifying every individual as a BPD case. The accuracy of this classification increased to around 72% for models containing between 3 and 9 predictors before decreasing with larger models. This method is highly promising, and provides evidence that a fairly simple blood-based gene expression screening of premature infants may aid in predicting their outcome and in selecting the proper course of treatment and monitoring for that newborn. Additional studies are underway using more sophisticated modeling techniques in an attempt to increase the predictive ability of this model.

5.5.0 Caveats

Significant genes identified within the Blommer et al. mouse lung development dataset were identified by ANOVA, while significant genes in the Kho et al. human lung development dataset were identified by principal component analysis. The use of two different selection criteria may introduce a bias in the comparison of genes identified in both datasets, as the genes have been identified by two different statistical methods. In order to combat this, one of the original datasets could be analyzed with the same method as the other, and the analysis could be performed once more.

The validation of genes of interest identified by differential expression analysis in PBMCs was performed in lung tissue. However informative these results may be in this case, in order to truly validate these genes, the analysis would ideally be performed using RNA samples from PBMCs. Also, all genes identified as significant within the PBMC data were increased in BPD, as the genes that were decreased did not meet the fold-change criteria. Therefore, there may be an opportunity for additional analyses to be performed with a less stringent threshold to examine genes that may also be decreased in the disease. The use of another platform, such as RNAseq, could also be valuable to these analyses by potentially increasing the sensitivity of the expression measurements.

There is a potential selection bias present within the results of the pathway analyses. There is a strong presence of pathways and GO classes that are linked with the inflammatory response. As PBMCs are largely comprised of inflammatory cells, it is highly likely that they would express genes linked with inflammatory pathways and classes. It is unclear whether these inflammatory pathways are truly representative of disease markers or if they have been identified as a result of performing the analyses with PBMC RNA. The use of specific cell-types rather than a group of multiple populations (PBMCs) could be used in order to normalize some of the expression values and eliminate the potential bias.
incurred by an overrepresentation or underrepresentation of one or more cell types in a given group of samples.

6.0.0 CONCLUSION

BPD is an extremely complex disorder consisting of multiple phenotypes contributed to by both genetic and environmental factors. It is the most common lung disorder of newborns, and yet little is known about the disease, especially the genetic component. This study provided a great deal of information about the genetic component using gene expression in PBMCs to identify genes that may be used as markers of BPD, and others that may also be linked with causal mechanisms of the disease. Utilizing multiple lung tissue datasets, the first specific aim of this study looked for genes of interest that may be linked with the causal mechanisms of BPD in lung tissue. The second examined peripheral blood mononuclear cells mainly for potential markers of BPD that may be helpful in the diagnosis and tracking of the disease state. However, causal mechanism and disease markers were not considered mutually exclusive to either of these studies. There is simply a higher likelihood of finding genes related to causal mechanism within the most directly affected tissue and disease markers in peripheral blood. Bronchopulmonary dysplasia is and will continue to be a major pediatric concern for the foreseeable future. These studies have provided valuable insight into the genetic aspects of bronchopulmonary dysplasia and may provide direction for future studies as well as potentially useful techniques like predictive modeling that could benefit both infants born prematurely as well as further research on the disease.
ACKNOWLEDGEMENTS

Advice on statistical modeling techniques and interpretation of results were given by Derick Peterson and Andrea Baran in the Department of Biostatistics at the University of Rochester.

qPCR samples were run with the help of Ashley Lopez and Valerie Langer of the Marian lab.

A sincere thank you to Mr. Soumyaroop Bhattacharya and Dr. Tom Marian for the time and effort that they put into every step of this project and furthering my education in general.

A final thank you to Dr. Michael Oster, Dr. Anne Haake, Mr. Soumyaroop Bhattacharya and Dr. Tom Marian for agreeing to be on this committee and for taking time out of their summer for the Defense.
8.0.0 WORKS CITED


2. Hayes, D; Frola, DJ; Murphy, BS; Shook, LA; Baxford, HC. Pathogenesis of Bronchopulmonary Dysplasia. Respir Care. 2010. 57:425-436.


34. Panczyszyn A; Wieczorek M. Role of CEACAM in neutrophil activation. Postepy Hig Med Dosw. 2012. 66:514-582.
36. Prsce JT; Ramalingam TR; Mentink-Kane MM; Wilson MS; El Kasmi KC; Smith AM; Thompson RW; Cheever AW; Murray PJ; Wynn TA. Arginase-1-expressing macrophages suppress Th2 cytokine-driven inflammation and fibrosis. PLoS Pathogens. 2009. 5(4):e1000371.
40. Clemmensen SN; Behr CT; Forvig S; Glemhticz A; Mora-Jensen H; Cramer EP; Jacobsen LC; Lassen MT; Cowland JB; Tanasi JI; Heegaard NH; Wren JD; Silhertahally AR; Borregaard N. Olfactomedin 4 defines a subset of human neutrophils. Journal of Leukocyte Biology. 2012. 91(3):495-500.
Appendix A: A-B Signal density histograms and hierarchical clusters for PLIER and RMA normalized near-birth samples, respectively. C-D Signal density histograms and hierarchical clusters for PLIER and RMA normalized near-diagnosis samples, respectively.