

Title: A systems genomics approach uncovers molecular associates of RSV severity

Authors: Matthew N. McCall^{1,2}, Chin-Yi Chu^{3,4†}, Lu Wang^{1†}, Lauren Benoodt⁵, Juilee Thakar^{1,6}, Anthony Corbett^{1,9}, Jeanne Holden-Wiltse^{1,9}, Christopher Slaunwhite^{3,4}, Alex Grier⁶, Steven R. Gill⁶, Ann R. Falsey^{7,10}, David J. Topham^{6,8}, Mary T. Caserta⁴, Edward E. Walsh^{7,10}, Xing Qiu^{1*} and Thomas J Mariani^{3,4*}

Affiliations:

¹Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester NY, USA.

²Department of Biomedical Genetics, University of Rochester Medical Center, Rochester NY, USA.

³Division of Neonatology and Pediatric Molecular and Personalized Medicine Program, University of Rochester Medical Center, Rochester NY, USA.

⁴Department of Pediatrics, University of Rochester Medical Center, Rochester NY, USA.

⁵Department of Biochemistry and Biophysics, University of Rochester Medical Center, Rochester NY, USA.

⁶Department of Microbiology and Immunology, University of Rochester Medical Center, Rochester NY, USA.

⁷Department of Medicine, University of Rochester Medical Center, Rochester NY, USA.

⁸David H. Smith Center for Vaccine Biology and Immunology, University of Rochester Medical Center, Rochester NY, USA.

⁹Clinical and Translational Science Institute, University of Rochester Medical Center, Rochester NY, USA.

¹⁰Department of Medicine, Rochester General Hospital, Rochester NY, USA.

*To whom correspondence should be addressed:

Xing Qiu; xing_qiu@urmc.rochester.edu

Thomas J Mariani: Tom_Mariani@urmc.rochester.edu

†These authors contributed equally to this work.

Abstract: Respiratory syncytial virus (RSV) infection results in millions of hospitalizations and thousands of deaths each year. Genetic variants resulting in alterations in the adaptive and innate immune response appear to be associated with RSV severity. To investigate the host response to RSV infection in infants, we performed a systems-level study of RSV pathophysiology, incorporating high-throughput measurements of the peripheral innate and adaptive immune systems and the airway epithelium and microbiota. We implemented a novel multi-omic data integration method based on multilayered principal component analysis, penalized regression, and feature weight back-propagation, which enabled us to identify cellular pathways associated with RSV severity. In both airway and immune cells, we found an association between RSV severity and activation of pathways controlling Th17 and acute phase response signaling, as well as inhibition of B cell receptor signaling. Dysregulation of both the humoral and mucosal response to RSV may play a critical role in determining illness severity.

[Main Text:]

Introduction

Respiratory syncytial virus (RSV), a negative strand RNA virus in the Pneumoviridae family, is a major cause of respiratory illness affecting persons of all ages, especially newborn infants¹⁻³. Although the majority of infections are relatively mild, RSV remains the most common cause of hospitalization for pneumonia and severe pneumonia in infants and young children in both the developed and developing world⁴⁻⁶. In the US half of newborns are infected in their first winter, with 1-3% hospitalized, 4-7% seen in emergency departments, and 10-16% seen in physician offices⁷.

A number of well-defined host factors predisposing infants to severe disease include prematurity, congenital cardiac and neuromuscular disease, and low levels of maternally derived neutralizing antibody³. More recent studies have also found genetic polymorphisms in cytokine and chemokine genes, altered innate interferon responses in the respiratory tract, T cell responses favoring a Th2 and Th17 bias, and the composition of the nasal microbiota to be associated with more severe illness⁸⁻¹⁵. Although each of these factors offer insight into the complex nature of RSV infection in young infants, they have generally been analyzed independently; thus, it is difficult to assess their interactions and relative importance to disease outcome.

Previous multi-omic analyses of RSV, by our group and others, demonstrated the potential of integrative analyses to further our understanding of the biological mechanisms underlying RSV disease progression and severity. To address limitations of prior studies, we designed a systems-level study of RSV pathophysiology in a precisely defined population of low risk newborns with the full spectrum of disease severity¹⁶. We studied purified populations of CD4+, CD8+ and CD19+ cells, as mixed PBMCs have been used extensively in studies of illness severity for respiratory infections^{9,11,17-19}. We reasoned that the local airway response would be a key component to defining illness severity, by contributing to the ability of the host to control or localize the infection²⁰. We also included data from the nasal microbiota, as recent studies have indicated colonization at the time of viral infection may significantly influence illness severity⁹.

This study builds on these previous studies both in the scope of the data and in the methodology developed to analyze these data. By modeling the connections between these high-throughput data and clinical RSV severity, we are able to reconstruct the intricate relationships among

different data types and demonstrate the potential of integrative analyses to identify shared and cell type specific cellular pathways associated with RSV severity.

Results

The data presented in this manuscript were generated as part of the Assessing Prediction of Infant Respiratory Syncytial Virus Effects and Severity (AsPIRES) study, which sought to identify host, viral, and environmental factors associated with RSV disease severity²¹. A total of 139 infants with RT-PCR confirmed RSV illness were enrolled. Venous blood, nasal microbiota and nasal epithelial cell samples were collected for high-throughput molecular analysis (Figure 1). Illness severity was measured using the Global Respiratory Severity Score (GRSS)²², which quantifies the full spectrum of primary RSV disease severity using nine clinical variables in a weighted score. We employ a novel approach to the integration and analysis of five high dimensional omic data modalities: the nasal epithelial transcriptome, the transcriptome of CD4, CD8, and CD19 cells from peripheral blood, and the nasal microbiome.

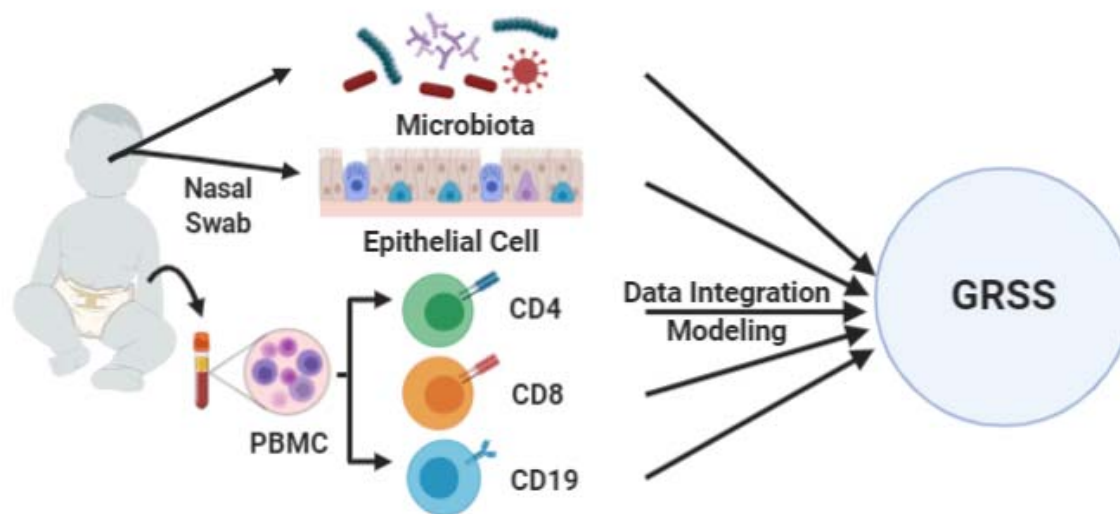


Figure 1: Overview of the study design. Measurements of the nasal epithelial cell transcriptome and nasal microbiome were generated from nasal swabs of infants. Measurements of peripheral immune cells (CD4, CD8, and CD19) were obtained from blood samples. These measurements were subsequently integrated and associated with RSV disease severity (GRSS).

Integrated Method Development

Preliminary exploratory analyses of individual data types found that most individual features (e.g. a single gene measured in CD4 cells) have relatively weak correlation with GRSS. Additionally, we observed strong correlation among features within a data type. These two observations motivated us to use appropriate dimension reduction methods to aggregate numerous “weak features” in the hopes of identifying a few key latent factors for each data type. Screeplots from performing PCA on each data type showed that a small number of PCs explained the vast majority of the variation in each data type (Supplementary Figure 1). Furthermore, we applied a secondary PCA on the PCs produced by each data type and found shared variation between data types. These initial analyses motivated the methods of data integration we describe in this paper.

Due to practical limitations, only a small subset of subjects (23 out of 106) have all five high dimensional data types (Supplementary Figure 2). Consequently, we decided to conduct integrative analyses of several combinations of omics data. The first set of data modalities was chosen to interrogate three disparate putative determinants of RSV severity, the nasal epithelial transcriptome (NT), the nasal microbiome (NM), and the adaptive immune response measured in peripheral blood CD4 cells (CD4). These data were available for 61 subjects. The second data integration focused on the collective role of the adaptive and innate immune response in RSV severity, measured in CD4, CD8, and CD19 cells isolated from peripheral blood. These data were available for 35 subjects.

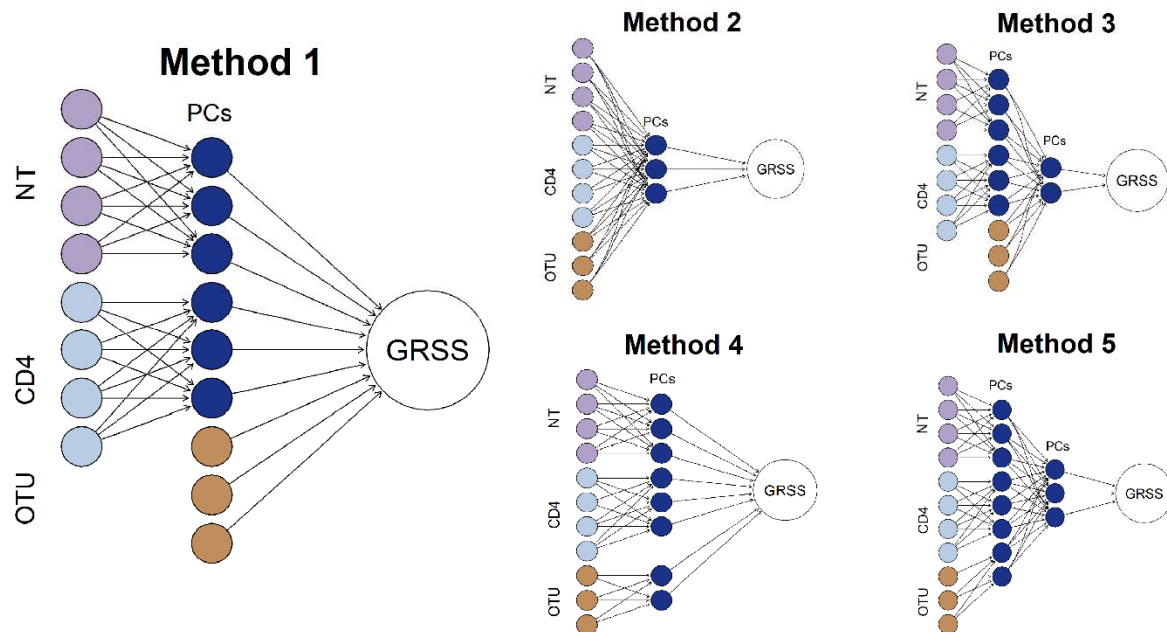


Figure 2: Five potential methods of multi-omic data integration. Methods differ in their handling of nasal microbiome data (OTU), number of levels of PCA, and the stage at which the integration occurs. In our assessments, the leftmost model out-performed the other models in terms of cross-validated error.

We proposed five related integrative analytic methods based on multilayer PCA and regularized regression based variable selection (Figure 2), see Methods for details. Based on extensive cross-validation (CV) experiments, we found that PCA of transcriptomic data, followed by an integrative elastic-net regression model with the transcriptomic PCs and individual operational taxonomic units (OTUs) of nasal microbiome data (Figure 2, Method 1) achieved better GRSS prediction accuracy than the other methods (Supplementary Tables 2 & 3).

We also applied a similar approach, initial PCA-based dimension reduction followed by regularized regression, to individual data types and found that the integrative models significantly out-performed the single data type models in terms of cross-validated prediction accuracy (Figure 3). Specifically, while the integrated and single data types models are all approximately unbiased, the integrated models had substantially smaller mean squared error (MSE).

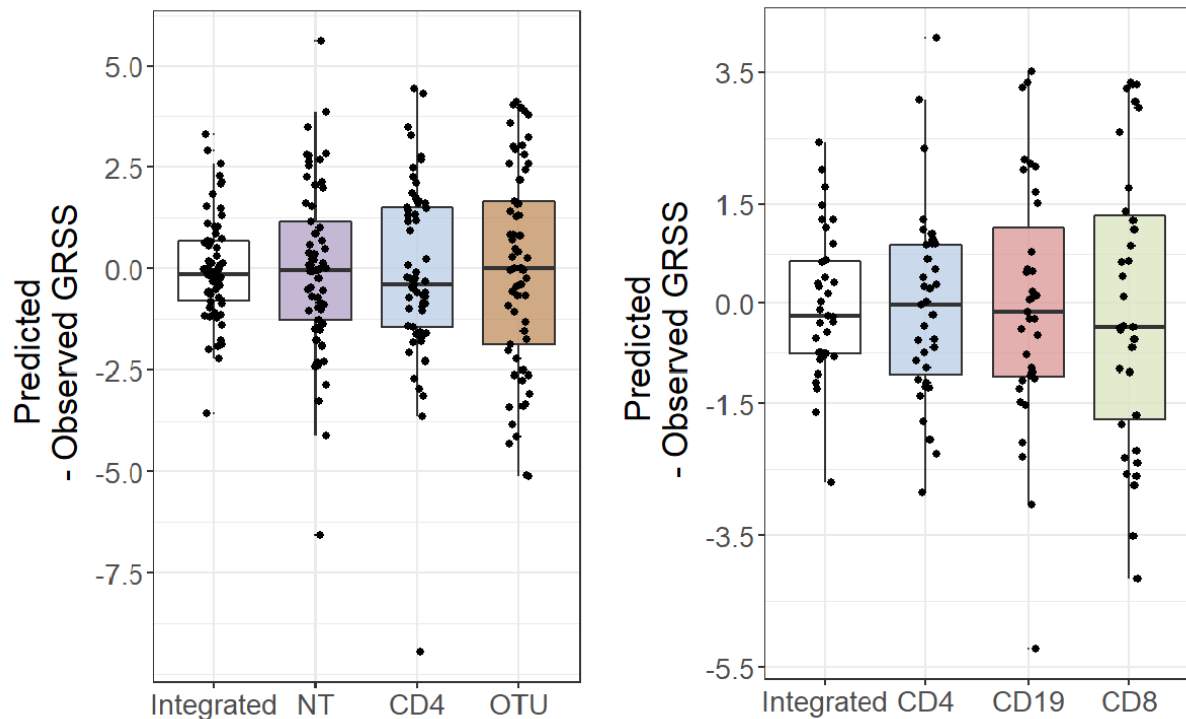


Figure 3: Integration of different data modalities improves prediction of GRSS. The panel on the left shows the difference between the predicted and observed GRSS for a model that integrates the nasal epithelial transcriptome, peripheral blood CD4 transcriptome, and nasal microbiome, as well as for models using just one of these data types. Similarly, the panel on the right compares a model that integrates the transcriptomes of 3 immune cell types measured in peripheral blood with models using just one of these data types. In both cases, integration increases the precision of the GRSS predictions.

Model Application and Interpretation

We previously reported gene expression correlates of clinical disease severity in RSV infected infants^{11,23}. To gain further insight, we generated multiple “models” by integrating unbiased, comprehensive gene expression data from both the humoral and mucosal compartments as described above. We hypothesized that a “systems level” analytical approach would provide distinct biological insights into disease pathophysiology. We first focused upon humoral responses specifically characterized in CD4 T cells sorted from peripheral blood collected during acute illness, in two models containing gene expression data from these cells (Figure 4). Following data integration, using RSV-associated disease severity as the outcome, the modeled weights (see “Feature weight calculation” in Supplementary Text for more details) for expression of individual genes is displayed in word clouds (upper left), and unweighted gene expression values are displayed in a heatmap (lower left). Weights for individual genes are clearly different between the two models, as evident from the word clouds, and could be expected due to the different subsets of the cohort included in each model. Interestingly, CD101, one of the highest weighted genes in both models, plays a role as an inhibitor of T cell proliferation induced by CD3. Furthermore, unweighted gene expression in these integrated models are not fundamentally different between mild and severe subjects. These observations support the novel and distinct insight derived from our new integrated modeling approach to identify gene expression

correlates of disease severity

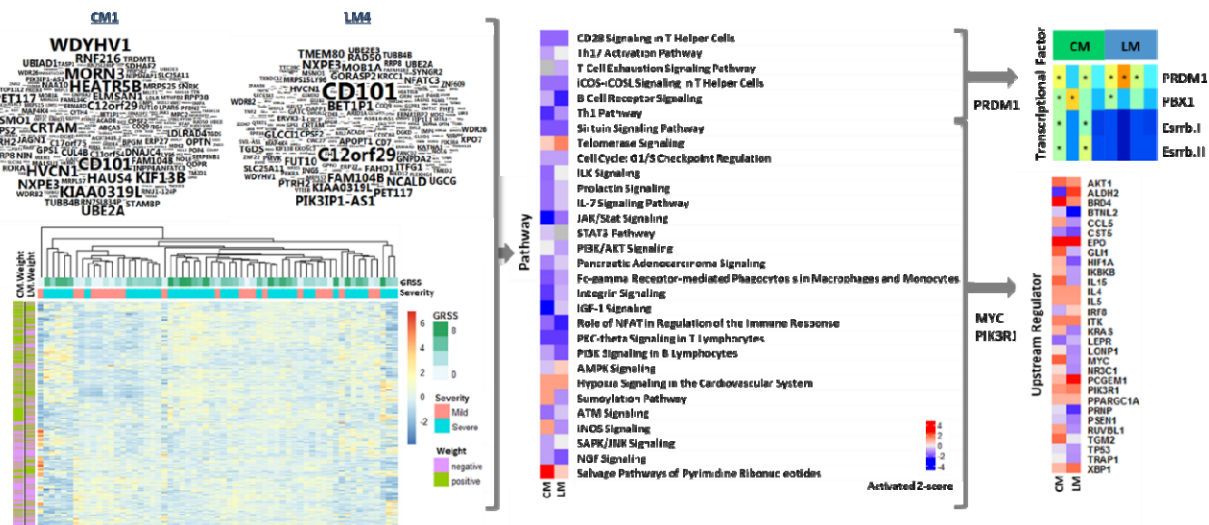


Figure 4: CD4 gene weights, expression, pathways, transcriptional factors and upstream regulators associated with clinical severity in integration models. Shown are an integration model of CD4, nasal epithelial cells and microbiota (CM) and a model of lymphocytes (LM). Weights generated by integration models are shown in word-clouds. The size of word represents the absolute value of gene weight. Word-clouds of CM & LM consist of genes that have absolute weight is greater than 0.0003. Gene expression are normalized expression levels for the 454 genes selected by univariate analysis; rows represent genes and columns represent samples. Red indicates higher expression, blue indicates low /no expression, green indicates Global RSV Severity Score (GRSS), soft orange indicates mild phenotype, lime green indicates severe phenotype, purple indicates negative weight and olive indicate positive weight. Transcriptional factors associated with severity in CD4 lymphocytes were identified using a hypergeometric test. Four transcriptional factors are shown where p-values were less than 0.05. Ingenuity Pathway Analysis (IPA) was used to identify canonical pathways and upstream regulators represented by genes associated with severity in CD4 lymphocytes. Thirty pathways and upstream regulators are shown where Fisher's exact test p-values were less than 0.05. Red and blue indicate predicted increased or decreased pathway activation (activation z-score), respectively.

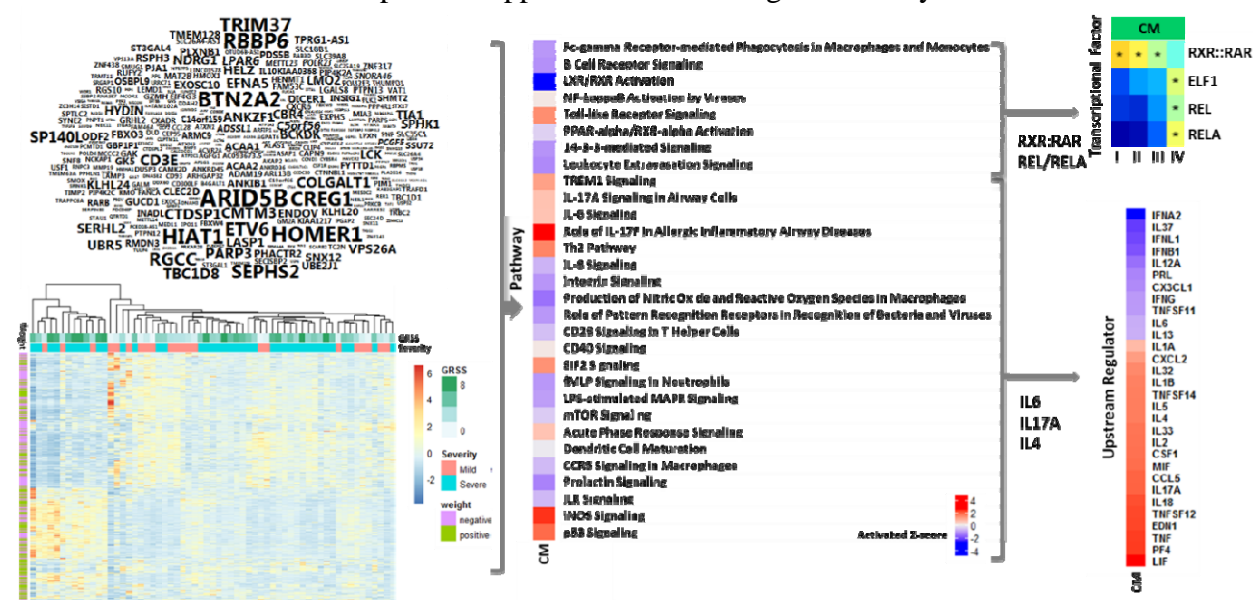
To minimize model-specific variation, we focused our interpretation of these data at the pathway level (middle of Figure 4 and Supplementary Figures 3-6). We found a high degree of model convergence, with a majority of pathways consistently, significantly activated or inhibited in association with disease severity. Significantly activated pathways were associated with hypoxia signaling, nucleotide salvage and telomerase signaling pathways. Of particular interest, many significantly inhibited pathways were associated with T helper (Th) cell signaling (CD28 and NFAT signaling) and differentiation of Th subtypes (Th1 and TH17 regulation).

We next attempted to predict key regulatory events associated with disease severity, including transcription factors and intracellular signal transduction molecules, that could be driving the global gene expression responses. Upstream regulator analysis using Ingenuity suggested that MYC and PI3K were associated with many of the significantly regulated pathways. Four sets of genes were selected based on their model weights: the top 200 genes, genes in the first quantile, genes in the second quantile, and all genes. Promoter analysis of these three sets were performed, using high quality transcription factor binding sites conserved across human, mice and rat genomes. The analysis revealed that multiple T cell differentiation pathways with consistent evidence for inhibition (CD28 signaling, Th17 activation, Th1 and T cell exhaustion signaling) are under the transcriptional control of PRDM1 (upper right/Panel D). Interestingly, PRDM1 is known to modulate peripheral T cell activation and proliferation, promote T helper (Th2) lineage commitment and limit Th1/Th17/Tfh cell differentiation. Our results implicate inhibition of T

cell differentiation, towards the Th2/Th17 phenotype in particular, as a putative mediator of severe illness.

We rationalized that the mucosal response would be distinct from the humoral response and would reflect the pathophysiology of the disease target organ, as suggested by our prior studies^{11,24}. Therefore, we next focused on interpreting the gene weights of nasal samples from the comprehensive model (CM). Similar to the CD4 data presented in Figure 4, the modeled weights for expression of individual genes in relation to disease severity is displayed in word clouds (Figure 5, upper left), and actual unweighted gene expression values are displayed in a heatmap (Figure 5, lower left). Highly weighted genes (positively correlated with disease severity) include *BTN2A2*, which inhibits the proliferation of CD4 and CD8 T-cells, T-cell metabolism and IFN secretion, and *HOMER1*, which negatively regulates T cell activation by inhibiting NFAT pathway.

We next completed pathway level analysis of the comprehensive model (CM) nasal gene expression weights (Figure 5, center). The analysis uncovered a decrease in multiple pathways driven by both retinoic acid-related (e.g., LXR/RXR and PPAR/RXR signaling) and p53-related (e.g., NF-kappaB) signaling were associated with severe disease. Further analysis of these genes and pathways identified significant evidence for changes in RXR and REL/A transcription factor regulation (Figure 5, upper right). Interestingly, this analysis indicated increased activation of pathways which were focused on regulation of the immune system were also associated with severe disease; in particular those associated with Th2 and Th17 CD4 T cells. Remarkably, upstream regulator analysis suggested this may be due to increases in the expression of *IL4* and *IL17A* (Figure 5, lower right). Again, as for CD4 T cell data, our ability to use this integrated modeling approach to identify evidence for pathophysiologically relevant interactions between the mucosal and humoral responses supports its methodological validity.



Transcriptional factors associated with severity in nasal epithelial cells were identified using a hypergeometric test. Four transcriptional factors are shown where *p*-values were less than 0.05. Ingenuity Pathway Analysis (IPA) was used to identify canonical pathways and upstream regulators represented by genes associated with severity in nasal epithelial cells. Thirty pathways and upstream regulators are shown where Fisher's exact test *p*-values were less than 0.05. Red and blue indicate predicted increased or decreased pathway activation (activation *z*-score), respectively.

Finally, we assessed unique and consistent responses in the mucosal and humoral compartments as indicated by our integrated models (CM and LM). Pathway-based analysis of the weights derived from CD8 T cells indicated unique activation of cytotoxic responses including those related to classical CD8 T cell functions were associated with severe disease. CD8 T cells also demonstrated unique activation of TGF β and TNFR signaling in severe disease. Conversely, analysis of the weights derived from CD19 B cells indicated regulation of multiple, alternate pathways. CD19 B cells displayed evidence for unique activation of PLC, and unique inhibition of PI3K/AKT signaling, among others. Pathway-based analysis of the weights also identified a number of responses that were conserved across all lymphocytes (e.g., CD4, CD8, CD19) and associated with disease severity. Consistently activated pathways indicated broad increases in oxidative phosphorylation, nucleotide salvage and sumoylation. Significantly inhibited pathways indicated broad reductions in lymphocyte proliferation, activation (e.g., iCOS and NFAT) and surtuin signaling. Finally, we looked for pathways which were consistently identified not just in lymphocytes, but across all humoral and mucosal data sets (Figure 6). This analysis indicated activation of pathways controlling Th17 and acute phase response signaling, as well as consistent inhibition of B cell receptor signaling, are consistently associated with disease severity in all cell types studied.

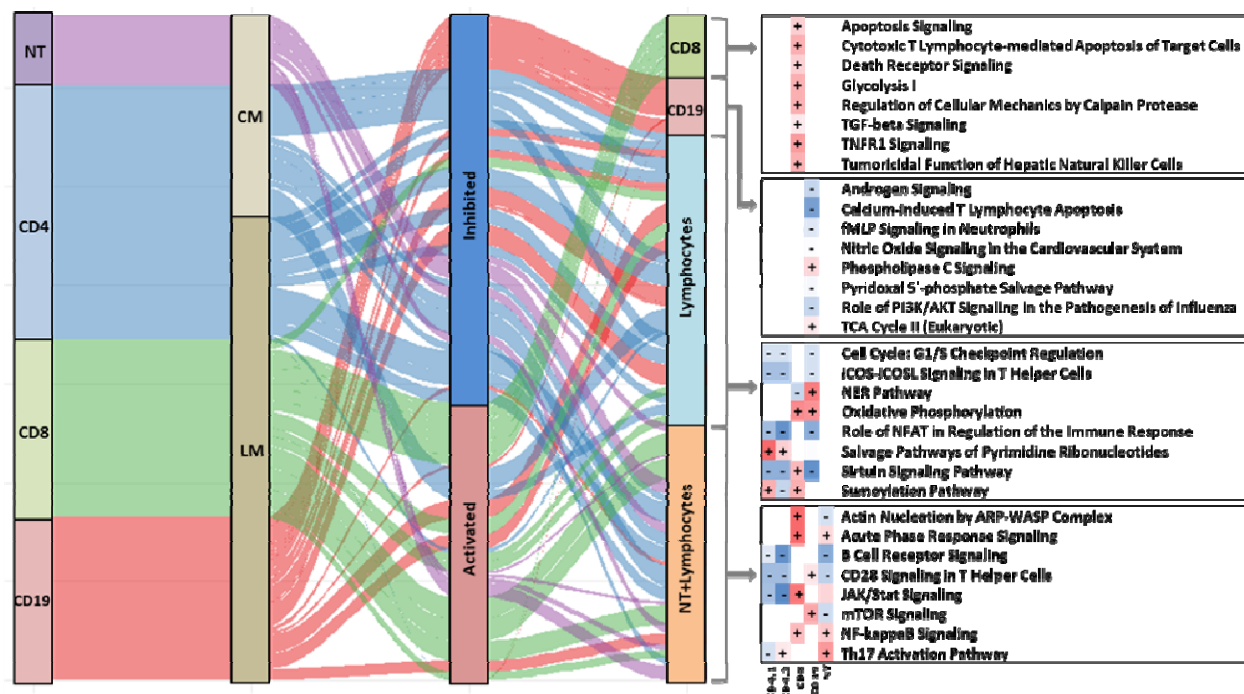


Figure 6: Common and unique pathways in lymphocytes & nasal epithelial cells of integration models. Sankey diagram showing the common and unique pathways among lymphocytes & nasal epithelial cells of integration model of CD4, nasal epithelial cells and microbiota (M1) and model of lymphocytes (M4). Ingenuity Pathway Analysis (IPA) was used to identify canonical pathways represented by genes associated with severity in lymphocytes and nasal epithelial cells. Pathways are shown where Fisher's exact test *p*-values were less than 0.05. Red and blue indicate predicted increased or decreased pathway activation (activation *z*-score), respectively. The width of the flow bar is proportional to absolute value of activation *z*-score.

Two alternative integration models were also considered in our study: one integrating CD4 and CD19 lymphocytes (Supplementary Figures 13-16) and another integrating nasal epithelial cells with CD4 and CD19 lymphocytes (Supplementary Figures 17-22). These additional models produced similar pathway level results for each transcriptomic data type, suggesting a degree of robustness to our approach to data integration and analysis and support our decision to focus on a comprehensive model of nasal epithelial cells, nasal microbiota, and CD4 lymphocytes (CM) and a model focused on three lymphocyte cell types (LM).

Discussion

In summary, we conducted a multi-omics study on infant RSV infection. We demonstrated that a multi-layer statistical learning framework was better at predicting disease severity than comparable single-layer approaches; and integrating multiple omics datasets provided with us better prediction accuracy of disease severity than predictors built from any single dataset. In addition, based on the trained sparse linear predictors, we were able to assess the contribution of individual genes/microbes by quantitative weights, which facilitate the biological interpretations of these predictive models.

Previous studies, including our own, utilized transcriptomic analysis of single sample types such as whole blood, purified T cells, or nasal secretions to investigate RSV disease pathogenesis. These analyses found an association between increased disease severity and both T and B cell suppression, evidence of systemic Th2 skewed T cell responses, alterations in systemic and local interferon responses, and the potential influence of the local microbiome on these responses. In the current work, by integrating gene expression data from multiple cell types from the peripheral blood and the respiratory epithelium with the complex microbiome of the upper respiratory tract, we confirmed the deregulation of the immune profile associated with RSV disease severity. Specifically, the results demonstrated that Th2 and Th17 activation, and inhibition of Th1 pathways dominate the T cell response. In addition, there was evidence of B cell suppression in the airway of infants with severe RSV. The results also demonstrated that inclusion of the microbiome, specifically *Haemophilus influenzae*, was informative for understanding a complete picture of RSV disease pathogenesis in young infants. Because our microbiome analysis did not identify *Streptococcus pneumoniae*, we were unable to confirm its influence. As described in the results, our analysis identified many severity-associated pathways that were activated or suppressed during infection, especially those indicating immune suppression. As shown in Figure 5, the airway cells contributed importantly to this inhibition. Airway epithelial cell expression identified BTN2A2 as an influential gene and was important in inhibiting CD4 and CD8 cells and interferon suppression.

Due to the high-throughput nature of omics data, there could be thousands of features that are potentially associative with biological endpoints. Besides, there exist an astronomically large number of possible interdependences among those omics data. Consequently, we believe that the most fundamental challenge in multi-omics studies such as ours is to reduce the complexity in statistical models in a sensible way, such that most spurious correlations are removed yet the major modes of informative relationships are retained.

Last but not the least, we believe our proposed method has better interpretability than those machine learning algorithms based on deep neural networks. In fact, the two methodologies share one architectural similarity, namely, both are multi-layered feedforward systems (see Fig.

2). Due to the deliberate choice of using linear activation function (via PCA) and linear output function (via glmnet regression), we were able to design a multilayer backpropagation algorithm that translates the weights at the output layer, i.e., linear coefficients in the trained predictor of GRSS, to mathematically equivalent weights at the input layer, i.e., individual genes and microbes. In a sense, the absolute value of a particular weight represents the contribution of this gene/microbe in predicting the GRSS. We performed gene set enrichment analyses based on these weights and discovered both cell-type specific and shared severity-associated biological pathways. In contrast, it is generally not possible to obtain such a direct relationship between the output and input layer in feedforward neural networks with nonlinear activation functions, which is why these algorithms are sometimes referred to as “blackbox methods”.

Arguably, several components in our integrative analysis pipeline may be replaced by other more specialized methods. For example, instead of using standard PCA, we could consider probabilistic PCA²⁵ or sparse PCA²⁶. The former is more robust to outliers and missing values due to the use of a ridge-like regularization term; and the latter is not only more consistent for “large p, small n” data than the standard PCA, it also has better interpretability because it can produce sparse loading scores. In addition, there is a host of recently developed variants of PCA that are advantageous in various situations, as summarized in²⁷.

Notwithstanding their advantages, these advanced dimension reduction methods are: (a) computationally more demanding; and (b) dependent on the tuning parameter that may have to be selected by cross-validation, which adds more computing cost and uncertainties in the analyses. In the future, we plan to systematically evaluate the impact of different dimension reduction strategies to the second stage integrative analyses and make algorithmic adaptations to improve their computational efficiency in high-throughput data analysis if necessary.

Another potentially rewarding future research direction is to design better statistical methods optimized for “incompletely matched” data like ours, in which all subjects do not have all types of data. In fact, as seen in Figure S2, only 23 subjects have all five types of high-throughput data, thus we decided to build integrative models for several combinations of data separately. Development of efficient and unbiased high-dimensional imputation methods in the future may allow us to fully integrate all available data and improve the accuracy of the predictive models.

What we present in this paper is a holistic approach to understanding the response to RSV infection during infancy, and the system-level correlates of clinical outcomes. The data provide a better resolution for accurately and sensitively identifying molecular changes associated with illness severity, and also uncover specific and robust changes that may be easily detectable using distinct biospecimens (e.g., airway, T cells, B cells, etc.). These novel data can guide future efforts to identify sensitive and specific biomarkers for identifying or predicting outcome following infant RSV infection. They may also be useful as biomarkers to inform the efficacy of future interventions (e.g., therapies) or preventative measures to suppress the rate of severe disease (e.g., vaccines). For example, our approach could potentially be used to quantify the response to novel RSV vaccines, either live attenuated or subunit vaccines, to be certain they do not mimic responses that may lead to a pathologic state^{28,29}.

Materials and Methods

Study population and sample collection

The full description of the AsPIRES study has been published²¹. The Research Subject Review Board approved the study and all families provided informed consent. Briefly, three groups of previously healthy full-term infants with primary RSV infection were evaluated during three winter seasons from 2012 to 2015 in Rochester, NY. RSV infection was identified in hospitalized infants, outpatients brought to medical attention either at the Emergency Department or primary care offices for respiratory symptoms, and infants in a birth cohort followed prospectively in their homes for RSV infection. From these three groups, infants with a range of RSV disease severity were included.

Infants were evaluated at three time points: an acute illness visit at diagnosis, a second visit ~14 days after illness onset, and a convalescent visit 28 days after illness onset. At each visit symptoms were recorded, and a physical exam was performed. At the first and third visit a flocked swab (Copan) was used to obtain a nasal specimen from one nares nares for microbiological testing, a nasal wash was performed on the contralateral nares to remove mucus and debris followed by the use of a second flocked swab to obtain epithelial cells by brushing the mucosa at the level of the inferior turbinate. Venous blood (~2-3 ml) was collected at each visit. RSV infections were confirmed by quantitative reverse transcriptase polymerase chain reaction (qRT-PCR) assay at diagnosis.

RSV disease severity was measured using the Global Respiratory Severity Score (GRSS), calculated using nine weighted clinical variables (general appearance; presence of wheezing, rales, retractions, cyanosis, lethargy, or poor air movement; maximal age-adjusted respiratory rate; and worst room-air oxygen saturation) yielding a score of 1 through 10²².

RNA processing

Four types of RNA-seq data were used in this study: NT (nasal transcriptome), CD4, CD8, and CD19. Technical details for recovering nasal RNA can be found in³⁰. Briefly, following flushing of the nares with saline to remove mucus and cellular debris, a flocked swab was used to recover cells at the level of the turbinates. The swab was immediately placed in RNA stabilizer (RNAprotect, Qiagen, Germantown, MD) and stored at 4 °C. Cells were recovered by filtering through a 0.45 µm membrane filter. Recovered cells were lysed and homogenized using the AbsolutelyRNA Miniprep kit (Agilent, Santa Clara, CA) according to the manufacturer's instructions.

CD4, CD8, and CD19 were mRNA expression profiles of the corresponding cell populations purified from peripheral blood as previously described^{11,31}. Specifically, heparinized blood was maintained at room temperature for up to 2 hours, and peripheral blood mononuclear cells were isolated by Ficoll-hypaque gradient, flow-sorted into these three subsets of cells, and stored in RNA lysis buffer at -20°C.

For all four types of RNA-seq data, sequencing libraries were constructed using the NexteraXT library kit (Illumina, San Diego, CA) and then sequenced on the Illumina HiSeq2500 platform. Sequences were aligned against human genome version of hg19 using STARv2.5, counted with HTSeq, and normalized by Fragments Per Kilobase of transcript per Million mapped reads (FPKM). A small subset of samples with very low yields or very low correlation with other samples were removed from the subsequent analyses, and we applied non-specific filtering based on both mean expression values and inter-quartile range (IQR) to identify subsets of genes for further investigation. We winsorized potential outliers at the gene-level, and then tested the correlation between gene expressions and the GRSS by Pearson correlation test. For each type of

transcriptomic data, we were able to select several hundreds of potentially informative features at 0.05 significance level, as summarized in Supplementary Table 1. Additional technical details on data preprocessing can be found in Supplementary Text.

Microbiome processing

Bacterial 16S rRNA from nasal swab specimens was extracted, amplified, and sequenced, and the resulting data were used to determine the taxonomic compositions, in terms of the relative abundances of those present operational taxonomic units (OTUs). Briefly, the V3-V4 hypervariable regions were targeted for amplification and sequenced using an Illumina MiSeq platform according to a paired end 2×300bp read protocol. Preliminary read processing and quality control were performed using the Quantitative Insights into Microbial Ecology (QIIME) software package^{32,33}, and a closed-reference OTU picking was done with USEARCH and the GreenGenes reference database³⁴. The initial microbiome data contained information for 148 distinct OTUs at the genus-level across 104 samples. Among them, only 15 genera had nonzero abundance level for more than half (n=52) of the subjects. These features were selected for the integrative analyses. A full list of them is provided in Supplementary Table 3.

Methods of data integration

We considered five methods of data integration, all of which shared three common components: a uni- or multi-layered dimension reduction to select a manageable set of features from various types of data, and elastic-net regularized regression to integrate these features into one weighted score to predict GRSS, our main outcome variable. Elastic-net regularized regression uses both L^1 (LASSO) and L^2 (ridge) penalties to produce a sparse linear predictive model and is known to be numerically stable for high-dimensional data. It is implemented by the R package glmnet³⁵. Regularization parameters were selected by an initial ten-fold cross-validation. After we obtained a sparse regression model, we re-estimated the linear coefficients by OLS-based procedures to improve the accuracy of modeling fitting. This parameter refinement strategy can improve predictive accuracy and is widely used in high-throughput data analysis³⁶.

Method 1 performed principal components analysis (PCA) separately on each transcriptomic data set and then uses the resulting PCs together with OTUs representing the nasal microbiota in a penalized regression model of GRSS. Methods 2 performed PCA collectively on all data types and then used the resulting PCs in a penalized regression model of GRSS. Method 3 was identical to Method 1 except for the addition of a second layer of PCA prior to a penalized regression model of GRSS. Method 4 was identical to Method 1 except the nasal microbiota OTUs also undergo dimension reduction via PCA. Finally, Method 5 combined the additions of Methods 3 & 4 resulting in a full two-layer PCA. We assessed the ability of the models produced by each of these methods to predict GRSS through leave-one-out cross-validation.

Weight Assignment and Transcription factor analysis

By design, the estimated linear coefficients in our integrative models are weights that represent the importance of principal components, not the features in the original data such as genes and

microbes. To enhance the interpretability of these integrative models, we calculated weights for each of the original features based on a backpropagation algorithm. Details of this calculation can be found in “Feature weight calculation” the Supplementary Text. These weights were then used in the transcription factor analyses. Conserved binding sites from JASPER across hg19, mm10 and rn6 were identified from as described previously³⁷. A hypergeometric test was performed to identify enriched binding sites^{38,39}.

Pathway analysis

Genes that were identified as significantly correlated with GRSS were subsequently used for canonical pathway identification and upstream regulator analysis using Ingenuity Pathway Analysis (QIAGEN Silicon Valley, Redwood City, CAQiagen). The combined feature weights were used in pathway analysis to enhance the results and were compared with the results obtained without those weights.

Data availability: Complete molecular and microbiota data for these studies is available in dbGaP (phs001201.v2.p1).

References and Notes:

- 1 Edwards, K. M. *et al.* Burden of human metapneumovirus infection in young children. *N Engl J Med* **368**, 633-643, doi:10.1056/NEJMoa1204630 (2013).
- 2 Falsey, A. R., Hennessey, P. A., Formica, M. A., Cox, C. & Walsh, E. E. Respiratory syncytial virus infection in elderly and high-risk adults. *N Engl J Med* **352**, 1749-1759, doi:10.1056/NEJMoa043951 (2005).
- 3 Meissner, H. C. Viral Bronchiolitis in Children. *N Engl J Med* **374**, 62-72, doi:10.1056/NEJMra1413456 (2016).
- 4 Causes of severe pneumonia requiring hospital admission in children without HIV infection from Africa and Asia: the PERCH multi-country case-control study. *Lancet* **394**, 757-779, doi:10.1016/s0140-6736(19)30721-4 (2019).
- 5 Jain, S. & Finelli, L. Community-acquired pneumonia among U.S. children. *N Engl J Med* **372**, 2167-2168, doi:10.1056/NEJMc1504028 (2015).
- 6 Shi, T. *et al.* Global, regional, and national disease burden estimates of acute lower respiratory infections due to respiratory syncytial virus in young children in 2015: a systematic review and modelling study. *Lancet* **390**, 946-958, doi:10.1016/s0140-6736(17)30938-8 (2017).
- 7 Hall, C. B. *et al.* Respiratory syncytial virus-associated hospitalizations among children less than 24 months of age. *Pediatrics* **132**, e341-348, doi:10.1542/peds.2013-0303 (2013).
- 8 Capella, C. *et al.* Prefusion F, Postfusion F, G Antibodies, and Disease Severity in Infants and Young Children With Acute Respiratory Syncytial Virus Infection. *J Infect Dis* **216**, 1398-1406, doi:10.1093/infdis/jix489 (2017).
- 9 de Steenhuijsen Piters, W. A. *et al.* Nasopharyngeal Microbiota, Host Transcriptome, and Disease Severity in Children with Respiratory Syncytial Virus Infection. *Am J Respir Crit Care Med* **194**, 1104-1115, doi:10.1164/rccm.201602-0220OC (2016).

- 10 Glezen, W. P., Paredes, A., Allison, J. E., Taber, L. H. & Frank, A. L. Risk of respiratory syncytial virus infection for infants from low-income families in relationship to age, sex, ethnic group, and maternal antibody level. *J Pediatr* **98**, 708-715, doi:10.1016/s0022-3476(81)80829-3 (1981).
- 11 Mariani, T. J. *et al.* Association of Dynamic Changes in the CD4 T-Cell Transcriptome With Disease Severity During Primary Respiratory Syncytial Virus Infection in Young Infants. *J Infect Dis* **216**, 1027-1037, doi:10.1093/infdis/jix400 (2017).
- 12 Miyairi, I. & DeVincenzo, J. P. Human genetic factors and respiratory syncytial virus disease severity. *Clin Microbiol Rev* **21**, 686-703, doi:10.1128/cmr.00017-08 (2008).
- 13 Sonawane, A. R. *et al.* Microbiome-Transcriptome Interactions Related to Severity of Respiratory Syncytial Virus Infection. *Sci Rep* **9**, 13824, doi:10.1038/s41598-019-50217-w (2019).
- 14 Thwaites, R. S. *et al.* Reduced Nasal Viral Load and IFN Responses in Infants with Respiratory Syncytial Virus Bronchiolitis and Respiratory Failure. *Am J Respir Crit Care Med* **198**, 1074-1084, doi:10.1164/rccm.201712-2567OC (2018).
- 15 Walsh, E. E. *et al.* Virus-Specific Antibody, Viral Load, and Disease Severity in Respiratory Syncytial Virus Infection. *J Infect Dis* **218**, 208-217, doi:10.1093/infdis/jiy106 (2018).
- 16 Chu, C.-Y. *et al.* Insufficiency in airway interferon activation defines clinical severity to infant RSV infection. *bioRxiv*, 641795, doi:10.1101/641795 (2019).
- 17 Bhattacharya, S. *et al.* Transcriptomic Biomarkers to Discriminate Bacterial from Nonbacterial Infection in Adults Hospitalized with Respiratory Illness. *Sci Rep* **7**, 6548, doi:10.1038/s41598-017-06738-3 (2017).
- 18 Bucasas, K. L. *et al.* Global gene expression profiling in infants with acute respiratory syncytial virus bronchiolitis demonstrates systemic activation of interferon signaling networks. *Pediatr Infect Dis J* **32**, e68-76, doi:10.1097/INF.0b013e318278b4b3 (2013).
- 19 Mejias, A. *et al.* Whole blood gene expression profiles to assess pathogenesis and disease severity in infants with respiratory syncytial virus infection. *PLoS Med* **10**, e1001549, doi:10.1371/journal.pmed.1001549 (2013).
- 20 Do, L. A. H. *et al.* Host Transcription Profile in Nasal Epithelium and Whole Blood of Hospitalized Children Under 2 Years of Age With Respiratory Syncytial Virus Infection. *J Infect Dis* **217**, 134-146, doi:10.1093/infdis/jix519 (2017).
- 21 Walsh, E. E. *et al.* Aims, Study Design, and Enrollment Results From the Assessing Predictors of Infant Respiratory Syncytial Virus Effects and Severity Study. *JMIR Res Protoc* **8**, e12907, doi:10.2196/12907 (2019).
- 22 Caserta, M. T. *et al.* Development of a Global Respiratory Severity Score for Respiratory Syncytial Virus Infection in Infants. *J Infect Dis* **215**, 750-756, doi:10.1093/infdis/jiw624 (2017).
- 23 Wang, L. *et al.* Airway Gene-Expression Classifiers for Respiratory Syncytial Virus (RSV) Disease Severity in Infants. *bioRxiv*, 628701, doi:10.1101/628701 (2020).
- 24 Chu, C. Y. *et al.* Airway gene expression correlates of RSV disease severity and microbiome composition in infants. *J Infect Dis*, doi:10.1093/infdis/jiaa576 (2020).
- 25 Tipping, M. E. & Bishop, C. M. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**, 611-622, doi:10.1111/1467-9868.00196 (1999).
- 26 Johnstone, I. M. & Lu, A. Y. On Consistency and Sparsity for Principal Components Analysis in High Dimensions. *J Am Stat Assoc* **104**, 682-693, doi:10.1198/jasa.2009.0121 (2009).
- 27 Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. *Philos Trans A Math Phys Eng Sci* **374**, 20150202, doi:10.1098/rsta.2015.0202 (2016).
- 28 Mejias, A., Rodríguez-Fernández, R., Oliva, S., Peeples, M. E. & Ramilo, O. The journey to a respiratory syncytial virus vaccine. *Ann Allergy Asthma Immunol* **125**, 36-46, doi:10.1016/j.anai.2020.03.017 (2020).

- 29 Vekemans, J. *et al.* Respiratory syncytial virus vaccine research and development: World Health Organization technological roadmap and preferred product characteristics. *Vaccine* **37**, 7394-7395, doi:10.1016/j.vaccine.2017.09.092 (2019).
- 30 Chu, C. Y. *et al.* The Healthy Infant Nasal Transcriptome: A Benchmark Study. *Sci Rep* **6**, 33994, doi:10.1038/srep33994 (2016).
- 31 Misra, R. S. *et al.* Flow-based sorting of neonatal lymphocyte populations for transcriptomics analysis. *J Immunol Methods* **437**, 13-20, doi:10.1016/j.jim.2016.07.001 (2016).
- 32 Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* **37**, 852-857, doi:10.1038/s41587-019-0209-9 (2019).
- 33 Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**, 335-336, doi:10.1038/nmeth.f.303 (2010).
- 34 DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**, 5069-5072, doi:10.1128/aem.03006-05 (2006).
- 35 Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* **33**, 1-22 (2010).
- 36 Belloni, A. & Chernozhukov, V. Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19**, 521-547, doi:10.3150/11-BEJ410 (2013).
- 37 Van Twisk, D., Murphy, S. P. & Thakar, J. Optimized logic rules reveal interferon- γ -induced modes regulated by histone deacetylases and protein tyrosine phosphatases. *Immunology* **151**, 71-80, doi:10.1111/imm.12707 (2017).
- 38 Thakar, J., Hartmann, B. M., Marjanovic, N., Sealfon, S. C. & Kleinstein, S. H. Comparative analysis of anti-viral transcriptomics reveals novel effects of influenza immune antagonism. *BMC Immunol* **16**, 46, doi:10.1186/s12865-015-0107-y (2015).
- 39 Zaslavsky, E. *et al.* Reconstruction of regulatory networks through temporal enrichment profiling and its application to H1N1 influenza viral infection. *BMC Bioinformatics* **14 Suppl 6**, S1, doi:10.1186/1471-2105-14-s6-s1 (2013).

Acknowledgments: The authors would like to thank the AsPIRES team for critical assistance with subject recruitment and sample collection and the UR Genomic Research Center for processing of the genomic samples. Finally, we are indebted to the patients and families who agreed to participate in these studies.

Funding: This project has been funded with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272201200005C, a University of Rochester School of Medicine and Dentistry Scientific Advisory Committee Incubator grant and University of Rochester Center for Clinical & Translational Science Institute grant number UL1 TR002001.

Author contributions: MNM, MTC, EEW, XQ, and TJM conceptualized the study. CC, SRG, MTC, EEW, and TJM designed the experiments. ARF, DJT, MTC, and EEW developed the cohort, and collected the specimens and clinical data. AC and JHW facilitated data organization, management and analysis. MNM, LW, LB, JT, and XQ developed the statistical and computational methods. MNM, CC, LW, LB, JT, CS, AG, SRG, MTC, EEW, XQ, and TJM generated, analyzed and interpreted the data. MNM, CC, LW, JT, SRG, MTC, EEW, XQ, and TJM wrote and/or revised the manuscript.

Competing interests: ARF is currently receiving funding from Merck Sharpe and Dohme, Pfizer, Janssen, Astra Zeneca and BioFire and personal fees for DSMB from Novavax. The others authors do not have any competing interests to report.