BMC Medical Genomics

RESEARCH ARTICLE Open Access

# Airway gene-expression classifiers for respiratory syncytial virus (RSV) disease severity in infants

Lu Wang[1†], Chin-Yi Chu[2†], Matthew N. McCall[1], Christopher Slaunwhite[2], Jeanne Holden-Wiltse[1], Anthony Corbett[1], Ann R. Falsey[3,5], David J. Topham[4], Mary T. Caserta[2], Thomas J. Mariani[2*], Edward E. Walsh[3,5*] and Xing Qiu[1*]

## Abstract

**Background:** A substantial number of infants infected with RSV develop severe symptoms requiring hospitalization. We currently lack accurate biomarkers that are associated with severe illness.

**Method:** We defined airway gene expression profiles based on RNA sequencing from nasal brush samples from 106 full-tem previously healthy RSV infected subjects during acute infection (day 1–10 of illness) and convalescence stage (day 28 of illness). All subjects were assigned a clinical illness severity score (GRSS). Using AIC-based model selection, we built a sparse linear correlate of GRSS based on 41 genes (NGSS1). We also built an alternate model based upon 13 genes associated with severe infection acutely but displaying stable expression over time (NGSS2).

**Results:** NGSS1 is strongly correlated with the disease severity, demonstrating a naïve correlation (ρ) of $\rho = 0.935$ and cross-validated correlation of 0.813. As a binary classifier (mild versus severe), NGSS1 correctly classifies disease severity in 89.6% of the subjects following cross-validation. NGSS2 has slightly less, but comparable, accuracy with a cross-validated correlation of 0.741 and classification accuracy of 84.0%.

**Conclusion:** Airway gene expression patterns, obtained following a minimally-invasive procedure, have potential utility for development of clinically useful biomarkers that correlate with disease severity in primary RSV infection.

**Keywords:** Respiratory syncytial virus, Respiratory severity score, Gene expression, RNA-seq, Classification

## Background

Respiratory syncytial virus (RSV) is the most important cause of respiratory illness in infants and young children, accounting for more than 57,000 bronchiolitis and pneumonia hospitalizations in the US annually [1]. Worldwide, 33.1 million acute lower respiratory infections and 3.2 million hospitalizations in children under 5 years of age are attributed to RSV each year [2]. In the US ~ 1–2% of newborns are hospitalized during their first winter, with rates greatest in the first two months of life (25.9 per 1000) [3]. Risk factors for severe disease include gestational age < 29 weeks, bronchopulmonary disease and symptomatic congenital cardiac disease, while less well defined risks include lack of breast feeding, and exposure to tobacco smoke. However, the majority of hospitalized infants are full-term infants whose only risk factor is young age at the time of infection [3].

*Correspondence: Tom_Mariani@urmc.rochester.edu; Edward.walsh@rochesterregional.org; xing_qiu@urmc.rochester.edu
†Lu Wang and Chin-Yi Chu contributed equally to this work
[1] Department of Biostatistics and Computational Biology, University of Rochester School Medicine, Rochester, NY, USA
[2] Department of Pediatrics, University of Rochester School Medicine, Rochester, NY, USA
[3] Department of Medicine, University of Rochester School Medicine, Rochester, NY, USA
Full list of author information is available at the end of the article

A number of severity scores using clinical parameters, including cutaneous oximetry, have been used to grade illness severity for use in management and as an outcome in therapeutic, or potentially, vaccine trials. [4–13] However, none of the clinically based severity scores have been universally adopted [14]. Reasons may include heterogeneity in the scope and purpose of the score, the ages to which it is applied and concerns about inter-observer variability and subjectivity in interpreting clinical signs, including oximetry, that often are temporally dynamic over short intervals. Identification of an objective biomarker that accurately correlates with, or potentially predicts, disease severity could be highly useful [15, 16].

We and others have reported a relationship between disease severity and host gene expression in peripheral blood cells and nasal swab samples during infection [17–20]. These results suggest such an approach may allow development of biomarkers to accurately categorize RSV disease severity. As part of the AsPIRES study [21] we recently reported on the feasibility of measuring gene expression of airway cells collected by nasal swab in healthy infants in order to study RSV disease pathogenesis [22]. However, in this manuscript, we describe the use of this gene expression data during RSV infection to develop two airway gene expression-based classifiers that are highly correlated with clinical disease severity. This represents a first step in developing a biomarker using gene expression responses capable of accurately classifying clinical severity in primary RSV-infection that could be used in conjunction with clinical evaluation.

## Methods

### Study subjects

Subjects included RSV infected infants enrolled in the AsPIRES study at the University of Rochester Medical Center and Rochester General Hospital [21]. RSV-infected infants came from three cohorts during three winters (October 2012 through April 2015); one cohort included infants hospitalized with RSV, a second cohort was recruited at birth and followed through their first winter for development of RSV infection, and the third cohort was RSV infected infants seen in pediatric offices and emergency departments and managed as outpatients. All subjects were full-term infants undergoing a primary RSV infection during their first winter season. Nasal samples were collected from the inferior nasal turbinate, by gentle brushing with a flocked swab as previously described [22], during the acute illness visit (visit 1) and at a convalescent visit ~ 28 after illness onset (visit 2). Illness severity was graded from 0 to 10 using a Global Respiratory Severity Score (GRSS), that uses nine parameters (age adjusted respiratory rate, chest retractions, wheezing, rales/rhonchi, apnea, cyanosis, room air oxygen saturation, lethargy and poor feeding) as previously described [23]. We defined a GRSS > 3.5 as severe disease as it is highly correlated with illness requiring hospitalization.

### Nasal RNA processing

The process for nasal RNA recovery was previously described [22]. Briefly, following flushing of the nares with 5 ml of 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, USA) and maintained at 4 °C. Cells were recovered by filtering through a 0.45 uM membrane filter. Recovered cells were lysed and homogenized using the AbsolutelyRNA Miniprep kit (Agilent, Santa Clara, CA, USA) according to the manufacturer's instructions. 1 ng of total RNA was amplified using the SMARter Ultra Low amplification kit (Clontech, Mountain View, CA, USA) and libraries were constructed using the NexteraXT library kit (Illumina, San Diego, CA, USA). Libraries were sequenced on the Illumina HiSeq2500. 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). This process was illustrated as Fig. 1 in [22]. As an alternative, we also tried replacing FPKM with CrossNorm [24–27], a normalization procedure designed for processing gene expression data with skewed patterns. In either case, a total of 6844 transcription profiles (genes) were reported after quality assurance analysis and preprocessing that include batch-effects removal and non-specific filtering. Additional technical details on data preprocessing can be found in Supplementary Methods and Supplementary Figure E1.

### Statistical methods

Descriptive statistics are reported in Table 1. Discrete variables are summarized in percentages, and continuous variables were summarized as Mean (STD). For continuous variables, we performed two-sample Welch t-tests to check the equality between the mild and severe groups; for categorical variables, Fisher's exact test was used instead. The nasal gene-expression severity scores we developed in this study were primarily based on multivariate regression analysis with bi-directional stepwise model selection based on Akaike Information Criterion (AIC). We also tried another model selection procedure based on elastic-net regularized regression, which uses both $L^1$ (LASSO) and $L^2$ (ridge) regression to produce a sparse regression model. The results are summarized in Supplementary Table E1. The R package glmnet [28, 29] was used for this purpose. Technical details of model

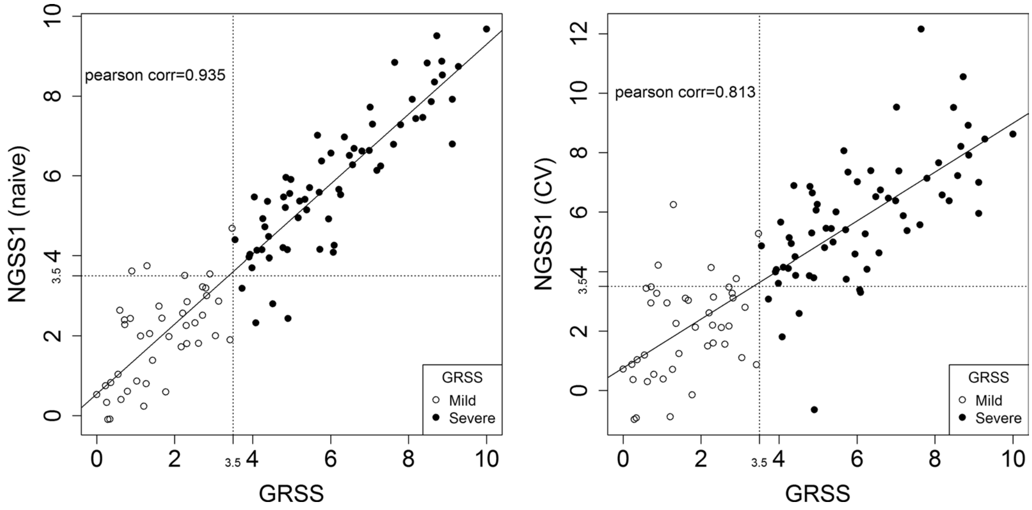

**Fig. 1** Correlating NGSS1 (severity score predicted by Model 2) with GRSS. Left: naïve Pearson correlation between GRSS and NGSS1 is $\rho = 0.935$. Right: cross-validated Pearson correlation between GRSS and NGSS is $\rho = 0.813$. Solid dots are subjects with severe symptoms (defined by GRSS > 3.5) and empty dots are those with mild symptoms (GRSS ≤ 3.5)

development and cross-validation (CV) can be found in section Model Developing and Cross-validation in Supplementary Methods. All analyses were conducted using SAS 9.3 (SAS Institute Inc., Cary, NC, USA) and the R programming language (version 3.5, R Foundation for Statistical Computing, Vienna, Austria).

## Results

Of the 139 RSV-infected infants enrolled in the AsPIRES study, nasal samples were available from 119 subjects during acute infection (day 1–10 of illness) and 81 subjects during convalescence (day 28 of illness). Among these 200 samples, 175 samples (106 acute samples and 69 convalescent samples) met sufficient quality to be used for subsequent analyses. Demographic and clinical information for these 106 subjects are provided in Table 1. The clinical severity score (GRSS) for these subjects ranged from 0 to 10, with 42 subjects considered to have mild disease (GRSS ≤ 3.5; mean ± SE GRSS of 1.63 ± 0.15) and 64 to have severe disease (GRSS > 3.5; mean GRSS of 6.13 ± 0.22). There were no significant differences between the mild and severe groups in gender, race, delivery type, breast feeding, or exposure to tobacco smoke. There also was no difference in age at time of infection or in duration of illness at the time of evaluation.

### Nasal gene expression correlates of clinical severity during acute illness

The 6,844 genes remaining after data preprocessing and filtering were subjected to the Pearson correlation test to select genes that were significantly correlated with GRSS during acute infection. After controlling the false discovery rate (FDR) at the 0.05 level, 66 significant genes were identified [30]. Using these genes, we applied model selection procedures (see Model Developing and Cross-validation in Supplementary Methods for more details) to select an initial multivariate regression model for GRSS (Model 1), which was comprised of 39 genes and had relatively good predictive power (77.4% accuracy, or 24 misclassifications) for the dichotomous clinical outcome (mild vs. severe illness) in leave-one-out cross-validation (LOOCV).

Not unexpectedly, there is a strong correlation among the 66 genes, which might reduce the diagnostic performance of Model 1. Using a novel method based on principal component analysis (PCA), we identify ten supplementary genes as additional features to model GRSS (see Model Developing and Cross-validation in Supplementary Methods for more details). With these additional features and using the same model selection strategy, we developed two additional models: Model 2 comprised of 41 genes and Model 3 comprised of 42 genes. The performance of these models was evaluated by LOOCV (Table 2). We found that the incorporation of the supplementary genes into Model 2 (CV prediction accuracy of 89.6%; 11 misclassifications) significantly improved the accuracy compared to Model 1 (24 misclassifications) and Model 3 (23 misclassifications). Of note, Model 2 contained 5 supplementary genes, and we defined it as NGSS1 (nasal gene expression severity score 1). As shown in Fig. 1, NGSS1 is highly associative with GRSS (naïve $\rho = 0.935$; CV $\rho = 0.813$). For the population of subjects in the AsPIRES study, the sensitivity and

**Table 1 Demographic data of subjects**

| | Mild (n = 42)[a] | | Severe (n = 64)[a] | | *p* value |
|---|---|---|---|---|---|
| | **n** | **Mean (STD) or %** | **n** | **Mean (STD) or %** | |
| Global Severity Score | 42 | 1.63 (1.00) | 64 | 6.13 (1.72) | $<$ 0.001 |
| Visit age (months) | 42 | 3.52 (1.99) | 64 | 3.24 (2.37) | 0.5122 |
| Gestational age (weeks) | 42 | 39.05 (1.25) | 64 | 38.8 (1.44) | 0.3437 |
| Birth weight (kg) | 42 | 3.32 (0.68) | 64 | 3.36 (0.57) | 0.7468 |
| Family size | 42 | 4.43 (2.86) | 64 | 3.98 (1.73) | 0.3703 |
| Days since disease onset | 42 | 4.31 (1.76) | 64 | 4.86 (1.78) | 0.1209 |
| Breast feeding summary | 42 | 1.56 (1.23) | 63 | 1.53 (1.25) | 0.8979 |
| Sex | 23 | 44.23 | 29 | 55.77 | 0.4275 |
| Male | | | | | |
| Female | 19 | 35.19 | 35 | 64.81 | |
| Ethnicity | 8 | 42.11 | 11 | 57.89 | 0.8018 |
| Hispanic or Latino | | | | | |
| Non-Hispanic or non-Latino | 34 | 39.08 | 53 | 60.92 | |
| Race | 23 | 37.1 | 39 | 62.9 | 0.3115 |
| Caucasian | | | | | |
| Other race | 19 | 47.5 | 21 | 52.5 | |
| Missing | - | 0 | 4 | 100 | |
| Delivery type | 29 | 36.71 | 50 | 63.29 | 0.3634 |
| Vaginal | | | | | |
| C-section | 13 | 48.15 | 14 | 51.85 | |
| Smoking exposure | 14 | 38.89 | 22 | 61.11 | 1 |
| Yes | | | | | |
| No | 28 | 40 | 42 | 60 | |
| RSV group | 23 | 38.98 | 36 | 61.02 | 1 |
| A | | | | | |
| B | 18 | 39.13 | 28 | 60.87 | |
| Missing | 1 | 100 | - | 0 | |

P-values reported in the last column were either based on Fisher's exact test (if the variable is categorical) or Welch *t*-test (if the variable is continuous). Continuous variables are reported as sample means (STD); categorical variables are reported as percentages

[a] Based on GRSS $\leq$ 3.5 (mild) or $>$ 3.5 (severe)

**Table 2 Performance of four models used in developing NGSS1 and NGSS2**

| | Number of genes selected | Naïve RSS | Naïve correlation | Naïve misclassified subjects (out of 106) | CV RSS | CV Correlation | CV prediction accuracy | CV misclassified subjects (out of 106) |
|---|---|---|---|---|---|---|---|---|
| Model 1 | 39 genes | 1.234 | 0.909 | 15 | 2.743 | 0.797 | 77.4% | 24 |
| Model 2* | **41genes** | **0.884** | **0.935** | **9** | **2.681** | **0.813** | **89.6%** | **11** |
| Model 3 | 42 genes | 0.920 | 0.933 | 13 | 2.119 | 0.844 | 78.3% | 23 |
| Model 4** | **13 genes** | **2.549** | **0.800** | **16** | **3.215** | **0.741** | **84.0%** | **17** |

Naïve and CV RSS are the mean residual sums of squares of the predictive model in the original and cross-validation analyses, respectively. Correlation are the Pearson correlation coefficient between the predicted severity scores and the clinically defined GRSS. Prediction accuracy is the percentage of correctly predicted mild (NGSS $\leq$ 3.5) or severe (NGSS $>$ 3.5) symptoms, compared with the same phenotype defined by the GRSS (mild: GRSS $\leq$ 3.5; severe: GRSS $>$ 3.5)

*Designated NGSS1. **Designated NGSS2

specificity for identifying severe disease were high (sensitivity 90.1%, specificity 88%) which would translate to a positive predictive value (PPV) of 92% and a negative predictive value (NPV) of 86%.

As a comparison, the Pearson correlation test identified 68 genes processed by CrossNorm with significant association with GRSS. The majority of them (39) were also detected by the FPKM normalized data, which showed that the two normalization methods are largely comparable. Using these 68 genes plus ten supplementary genes identified by PCA as candidate genomic features, we developed Model 2b, which had similar but slightly worse prediction accuracy in LOOCV experiments. Technical details of the development of Model 2b and the summary of its performance are provided in An Alternative Method Based on CrossNorm in Supplementary Methods and Supplementary Table E4. We will focus on FPKM normalized data from now on.

### Validation of NGSS1 at the convalescence phase

NGSS1 was trained exclusively from data collected at the acute phase (visit 1). For a subset (n = 54) of subjects, we also had their nasal transcriptome profiles at the convalescence phase (day 28 after illness onset), a time when most infants had completely recovered from their illness. If NGSS1 is a valid surrogate for disease severity, we hypothesized that NGSS1 calculated from the severely ill subjects at visit 2 would converge to those of the mildly ill subjects. Compared with the acute visit, the calculated NGSS1 at the convalescent visit predicted a significantly lower mean severity score for severe subjects (n = 29, 6.22 vs. 2.82, $p < 0.001$). In contrast, there was no significant difference in NGSS1 between the two visits for the mildly ill group (n = 25, 1.96 vs. 2.31, $p = 0.45$), nor between the severe and mild groups at visit 2 (2.82 vs. 2.31, p = 0.40). These results are illustrated in Fig. 2a.

### Exploratory association analysis based on stable nasal genes

In the process of developing NGSS1 we observed that a large number of genes had expression levels that remained stable between the acute and convalescent visits. We speculated that a NGSS based on stable genes that were correlated with GRSS could potentially be predictive of disease severity prior to illness onset. Thus, we next developed an NGSS based on genes displaying stable expression across acute illness and convalescence in the 54 subjects with samples from both time points. Specifically, we included only genes whose mean expression levels correlated with disease severity during acute illness, and whose expression did not change significantly from the acute to convalescent stage.

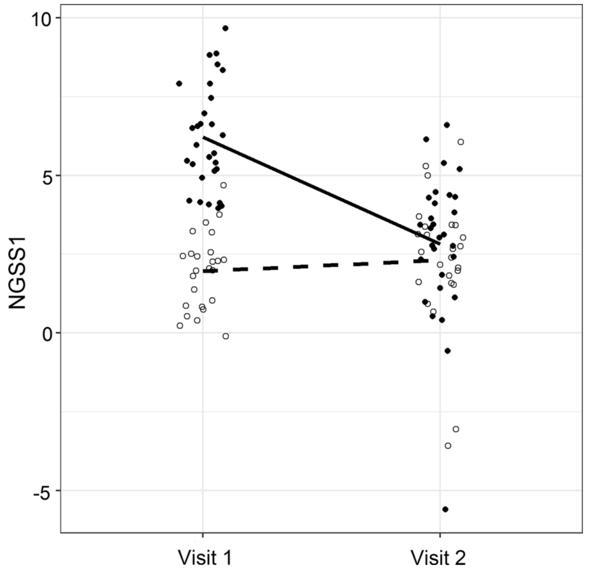

**Fig. 2** Paired comparisons between visit 1 and visit 2 using NGSS1 (panel **a**) and NGSS2 (panel **b**). A total of n = 54 subjects with samples in both visits were used. Solid dots represent severe subjects and empty dots represent mild subjects. The solid line represents the mean trend of severe subjects and the broken line represents the mean trend for mild subjects. **a** At visit 1, there was a significant difference in mean NGSS1 between the severe (n = 29) and mild (n = 25) groups (6.22 vs. 1.96, $p < 0.001$). Mean NGSS1 of the mild group was virtually unchanged between two visits (1.96 vs. 2.31, $p = 0.45$). In comparison, mean NGSS1 of the severe group declined significantly at visit 2 (6.22 vs. 2.82, $p < .001$). **b** In contrast to NGSS1, the differences in NGSS2 was virtually unchanged between the two visits, due to the fact that NGSS2 were built with stable genes

We identified 2127 genes in subjects with mild illness and 1531 genes in subjects with severe illness, based on paired two sample t-test ($p > 0.5$) and fold change increases or decreases within 10%. Of the total 3658 genes, 689 stable genes were common in both groups (Fig. 3a). A quality assurance analysis based on IQR showed that a small subset (n = 14) of these genes had relatively small dynamic range in the combined dataset, and were excluded. We applied marginal screening based on Pearson correlation with GRSS to the remaining 675 stable genes and identified 44 marginally significant genes. As in developing NGSS1, we added 5 supplementary genes with strong marginal associations with GRSS. Model selection identified 13 genes as Model 4 (designated as NGSS2). The performance of NGSS2 is provided in Table 2 and illustrated in Fig. 3b. NGSS2 showed a significant correlation with GRSS (ρ = 0.741), and a CV accuracy of 84% (17 misclassifications out of 106 cases, Table 2). Of note, NGSS1 and NGSS2 do not contain any commonly selected gene, which is expected due to different screening criteria. Figure 2b shows that on average, NGSS2 did not change between visit 1 and visit 2, which

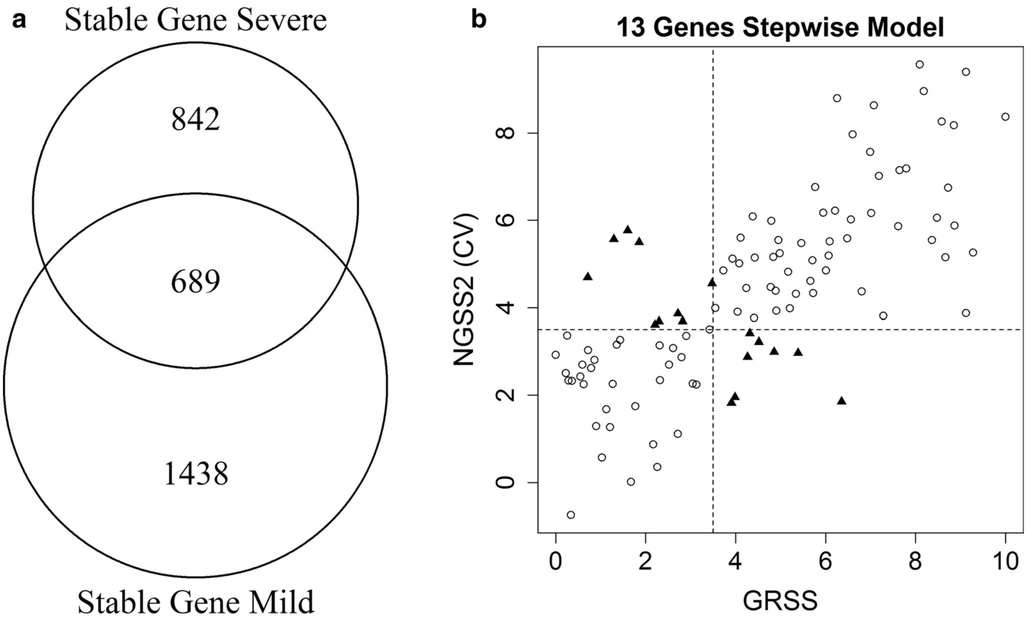

**Fig. 3** **a** Diagram indicating the stable genes for the mild (GRSS ≤ 3.5) and severe (GRSS > 3.5) groups and the 689 intersecting stable genes common to both groups. **b** Correlating NGSS2 (severity score predicted by Model 4) with GRSS. Naïve Pearson correlation between GRSS and NGSS2 is $\rho = 0.800$. Right: cross-validated Pearson correlation between GRSS and NGSS is $\rho = 0.741$. Circles are subjects with correct cross-validated classification based on NGSS2; solid triangles are misclassified subjects

is the key difference between these two classifiers. A full list of genes used in NGSS1 and NGSS2, as well as their estimated linear coefficients in the models, are listed in Supplementary Tables E2 and E3.

## Discussion

Several approaches have been proposed for quantifying RSV disease severity in young infants [4–13]. A variety of clinical parameters have been included in several described severity scores, with incomplete agreement on the optimal factors to select [14]. One reason is that many clinical signs of RSV infection in young infants, including cutaneous oximetry, can fluctuate frequently and rapidly during the course of illness, making consistent assessment difficult. In fact, even the direct measurement of RSV viral load in respiratory secretions is not significantly correlated with disease severity in the AsPIRES study [31]—similar phenomenon was also reported by several other similar studies [32–34]. An objective biomarker reliably correlated with clinical severity could prove useful for clinical management and as a classifier and/or an outcome measure in vaccine or therapeutic trials.

Transcriptomic analysis of host cells has proven informative in the study of several respiratory viral infections, including RSV, with the emphasis on disease pathogenesis [17–20]. Unlike this report that focuses on nasal epithelial cell samples, most reports have described gene expression correlates of disease severity in peripheral blood mononuclear cells during infection since RSV pathogenesis is thought to be closely linked to the host's immune response [35]. In two publications from the same group, RSV infection was associated with overexpression of innate immunity genes (neutrophil and interferon genes) and suppression of adaptive T and B cell genes. [17, 19] The investigators used the results to develop a gene-expression based illness score (designated Molecular Distance to Health [MDTH]) that was significantly correlated with a clinical disease severity score, duration of hospitalization and need for supplemental oxygen. Recently, Jong et al. described an 84 gene signature that was highly predictive of RSV disease severity in infants [16]. Similarly, we reported that gene expression patterns in purified blood CD4 T cells during infection were correlated with clinical disease severity [18]. Gene expression results from nasal swabs collected from hospitalized infants during RSV infection have also been recently reported by another group, with differentially expressed genes correlated with clinical severity [20].

Although the nasal brush samples from the AsPIRES study were collected to investigate molecular pathways and disease mechanisms involved in pathogenesis (presented in a separate manuscript [36]), we also considered that the data could be useful for the development of a gene based biomarker of RSV severity. We used marginal screening of all genes followed by PCA analysis and step-wise model selection to develop NGSS1, a multivariate linear classifier of severity. In CV analysis, NGSS1 was strongly correlated with GRSS and was a relatively accurate classifier of binary disease severity. Furthermore, the score tracked well with clinical improvement 28 days after illness onset. Of particular note, we found that including uncorrelated supplementary genes enhances the accuracy of the models, and recommend this approach as a routine for future classification/prediction analyses based on high-throughput data with substantial correlation. Another point worth mentioning is that the data went through a thorough pre-processing prior to model development. This "defensive preprocessing" not only greatly reduced aberrations in the data, but also guaranteed that the selected candidate features were highly informative. Consequently, the particular choice of normalization procedure (FPKM or Cross-Norm) became less important. As noted, in the population enrolled in our study the operating characteristics of NGSS1, including sensitivity, specificity, PPV and NPV, were quite good. However, it should be recognized that the proportion of mildly ill to severely ill subjects was determined by the recruiting strategy used, and that the PPV and NPV would vary depending on the population to which NGSS1 was applied [21]. If mildly ill subjects are increased by a factor of 3–5 this would reduce the PPV to 40–70% although the NPV would remain > 90%.

Although the aim of this report is not to describe molecular mechanisms operative during RSV infection, it should be noted that the 41 NGSS1 genes include cytokines (TNFSF10, IL6, and CXCL2), extracellular matrix proteins (VIM, MMP19, RPS15A, FKBP1A, and VCAN), inflammation regulators (CXCL2, CD163), and components of various signaling processes (GNS, HAVCR2, PTPRC, CTSL, INHBA, IL6, MMP19, CXCL2, SLC39A8, CCDC80, VCAN, CD163). Some genes are only known to be involved in fundamental biological processes and are therefore novel in RSV research, including ST3GAL1 (a type II membrane protein) and ATP10B (ATPase Phospholipid Transporting 10B). Note that only two genes (TNFSF10, RABGAP1L) have been associated with disease severity in our recent study based on purified CD4 T cells [18]. In addition, IL-6 Signaling is the only significant canonical pathway identified from the CD4 T cells that contains an NGSS1 gene (IL6).

A unique and very preliminary result from our analysis is the development of NGSS2 using differentially expressed genes associated with GRSS that did not change between the acute and the convalescent time points. It is possible that these genes may simply be slow to return to baseline expression levels, in contrast to those genes selected for NGSS1. Although speculative, it occurred to us that "stable" genes might possibly be predictive of severity regardless of when a nasal sample was obtained, thus raising the possibility of infants at risk prior to or early in infection. While NGSS2 is slightly less accurate than NGSS1 in predicting GRSS during acute illness, the association between NGSS2 and GRSS is still relatively strong. Interestingly, the 13 NGSS2 genes were broadly related to cytoplasmic activities (EXOSC10, PLK2, PPIC, CLDN10, MAP3K13, MT1G, PXN), ATP binding (SEPHS2) and phosphoprotein regulation (BCKDK, PLK2, MAP3K13); activities that may be less directly responsive to acute RSV infection. These observations suggest that the best nasal transcriptome predictors of respiratory symptoms are not necessarily limited to those genes that directly regulate the immune response to RSV infection.

The use of nasal brush specimens for development of a severity biomarker in infants is attractive for a number of reasons. Nasal respiratory epithelial cells are the first cells infected and directly initiate early innate immune responses to RSV. The mucosa is also the site of migration of both innate and adaptive immune cells during infection. Importantly, we have shown that gene expression in nasal respiratory epithelial cells is highly concordant with published gene expression in lower respiratory tract epithelial cells, and thus should be a reasonable proxy for lung responses to RSV infection [22]. Of practical importance, collection of nasal epithelial cells is relatively non-invasive and simple to perform with minimal discomfort.

There are several important limitations to our study and conclusions. First, we do not have an independent cohort to validate our findings; the only publically available nasal gene expression data during RSV infection used microarray technology that did not identify many of the genes we identified by RNAseq. Due to the lack of independent samples for validation, we applied cross-validation techniques to prevent model overfitting and validate the accuracy of prediction for both NGSS1 and NGSS2 at the acute visit. CV estimator for prediction accuracy is known to be asymptotically unbiased [37] under very weak statistical assumptions, namely, the training and testing data are independent and identically distributed (which can even be relaxed further, see [38, 39]). Additionally, we further validated the NGSS1 trained at the acute visit with the convalescence data, and the results conformed with our prediction remarkably well. Note that the acute infection and convalescence visits are reasonably spaced out (about 2 ~ 3 weeks apart), therefore they are nearly independent: the mean and median serial correlations among all filtered genes is 0.103 and 0.095, respectively. Although the NGSS1 declined for the severely ill infants when clinical symptoms had resolved, it would be useful to determine if NGSS1 tracked closely over the full course of an illness.

However, validation of our findings with an independent prospective cohort will be required. In addition, the results may not be valid for infants older than 10 months of age when infected with RSV, nor for infants with prematurity or other underlying medical conditions.

Another possible limitation is that all data used in these analyses were generated on the same technical platform and processed by the same team, therefore the validation results do not reflect the impact of "artifacts" in transcriptomic studies such as batch effects and platform differences, which can be reduced but not entirely eradicated by advanced normalization methods.[40–43]

Importantly, speculation that NGSS2 might predict disease severity prior to infection demands careful prospective validation. Finally, to extend the utility of time-intensive gene expression assays beyond a research tool and use it as a clinically useful biomarker of RSV disease severity, will require translation of these results to a rapid readily performed multiplex reverse transcription polymerase chain reaction (RT-PCR) assay, similar to those that have recently been developed for microbial diagnostics in respiratory secretions [44].

## Conclusions

In this report, we demonstrate that gene expression data obtained from an easily and safely obtained nasal brush specimen in young infants with acute RSV infection shows promise for development of composite molecular biomarkers that closely correlate with clinical severity score. Using a statistical learning procedure based on marginal screening, PCA, and multiple regression with stepwise model selection, we developed two nasal gene-expression severity scores (NGSS1 and NGSS2) that are highly correlated with a clinically derived disease severity score (GRSS). Further studies to refine and validate the potential of predictive gene expression data from readily collected nasal samples are needed.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12920-021-00913-2.

**Additional file 1:** Additional technical details of RNA-Seq pre-processing, predictive model developing and cross-validation, full lists of genes selected in NGSS1 and NGSS2, and some supplemental results related to CrossNorm.

### Abbreviations

AIC: Akaike information criterion; CV: Cross-validation; FDR: False discovery rate; FPKM: Fragments per kilobase of transcript per million mapped reads; GRSS: Global respiratory severity score; IQR: Inter-quartile range; LOOCV: Leave-one-out cross-validation; NGSS1: Nasal gene expression severity score 1; NGSS2: Nasal gene expression severity score 2; NPV: Negative predictive value; PCA: Principal component analysis; PPV: Positive predictive value; RSV: Respiratory syncytial virus.

### Acknowledgements

An abstract of this study was accepted in *Open Forum Infectious Diseases*, 2019 vol. 6, pp. S210-S210.

### Authors' contributions

XQ, TJM, and EEW conceptualized the study. TJM, EEW, MTC, and CC designed the experiments. EEW, MTC, ARF and DJT developed the cohort, and collected the specimens and clinical data. LW, MNM, and XQ developed statistical models. JHW and AC facilitated data organization, management and analysis. LW, CC, MNM, CS, JH-W, AC, ARF, DJT, MTC, TJM, EEW, and XQ generated, analyzed and interpreted the data. LW, CC, MNM, CS, JH-W, AC, ARF, DJT, MTC, TJM, EEW, and XQ wrote and/or revised the manuscript. All authors read and approve the final manuscript.

### Funding

This study is supported in part by Respiratory Pathogens Research Center (NIAID contract number HHSN272201200005C), and the University of Rochester CTSA award number UL1 TR002001 from the National Center for Advancing Translational Sciences of the National Institutes of Health. The University of Rochester Genomics Research Center supported generation of the sequencing data. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

### Availability of data and materials

The transcriptional data described in this manuscript are available in dbGaP (phs001201.v2.p1).

### Ethics approval and consent to participate

This study was approved by the Institutional Review Boards of the University of Rochester Medical Center and Rochester General Hospital. All parents provided written informed consent.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

### Author details

[1] Department of Biostatistics and Computational Biology, University of Rochester School Medicine, Rochester, NY, USA. [2] Department of Pediatrics, University of Rochester School Medicine, Rochester, NY, USA. [3] Department of Medicine, University of Rochester School Medicine, Rochester, NY, USA. [4] Department of Microbiology and Immunology, University of Rochester School Medicine, Rochester, NY, USA. [5] Department of Medicine, Rochester General Hospital, Rochester, NY, USA.

Received: 20 August 2020 Accepted: 19 February 2021
Published online: 25 February 2021

### References

1. Hall CB, Weinberg GA, Iwane MK, Blumkin AK, Edwards KM, Staat MA, Auinger P, Griffin MR, Poehling KA, Erdman D, *et al*. The burden of respiratory syncytial virus infection in young children. N Engl J Med. 2009;360(6):588–98.
2. Shi T, McAllister DA, O'Brien KL, Simoes EAF, Madhi SA, Gessner BD, Polack FP, Balsells E, Acacio S, Aguayo C, *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. 2017;390(10098):946–58.
3. Hall CB, Weinberg GA, Blumkin AK, Edwards KM, Staat MA, Schultz AF, Poehling KA, Szilagyi PG, Griffin MR, Williams JV, *et al*. Respiratory syncytial virus-associated hospitalizations among children less than 24 months of age. Pediatrics. 2013;132(2):e341-348.
4. Bekhof J, Reimink R, Brand PL. Systematic review: insufficient validation of clinical scores for the assessment of acute dyspnoea in wheezing children. Paediatr Respir Rev. 2014;15(1):98–112.
5. Corneli HM, Zorc JJ, Holubkov R, Bregstein JS, Brown KM, Mahajan P, Kuppermann N. Bronchiolitis Study Group for the Pediatric Emergency Care

Applied Research N: Bronchiolitis: clinical characteristics associated with hospitalization and length of stay. Pediatr Emerg Care. 2012;28(2):99–103.

6. Destino L, Weisgerber MC, Soung P, Bakalarski D, Yan K, Rehborg R, Wagner DR, Gorelick MH, Simpson P. Validity of respiratory scores in bronchiolitis. Hosp Pediatr. 2012;2(4):202–9.
7. Duarte-Dorado DM, Madero-Orostegui DS, Rodriguez-Martinez CE, Nino G. Validation of a scale to assess the severity of bronchiolitis in a population of hospitalized infants. J Asthma. 2013;50(10):1056–61.
8. Feldman AS, Hartert TV, Gebretsadik T, Carroll KN, Minton PA, Woodward KB, Larkin EK, Miller EK, Valet RS. Respiratory severity score separates upper versus lower respiratory tract infections and predicts measures of disease severity. Pediatr Allergy Immunol Pulmonol. 2015;28(2):117–20.
9. Gajdos V, Beydon N, Bommenel L, Pellegrino B, de Pontual L, Bailleux S, Labrune P, Bouyer J. Inter-observer agreement between physicians, nurses, and respiratory therapists for respiratory clinical evaluation in bronchiolitis. Pediatr Pulmonol. 2009;44(8):754–62.
10. McCallum GB, Morris PS, Wilson CC, Versteegh LA, Ward LM, Chatfield MD, Chang AB. Severity scoring systems: are they internally valid, reliable and predictive of oxygen use in children with acute bronchiolitis? Pediatr Pulmonol. 2013;48(8):797–803.
11. Mosalli R, Abdul Moez AM, Janish M, Paes B. Value of a risk scoring tool to predict respiratory syncytial virus disease severity and need for hospitalization in term infants. J Med Virol. 2015;87(8):1285–91.
12. Parker MJ, Allen U, Stephens D, Lalani A, Schuh S. Predictors of major intervention in infants with bronchiolitis. Pediatr Pulmonol. 2009;44(4):358–63.
13. Fernandes RM, Plint AC, Terwee CB, Sampaio C, Klassen TP, Offringa M, van der Lee JH. Validity of bronchiolitis outcome measures. Pediatrics. 2015;135(6):e1399-1408.
14. Karron RA, Zar HJ. Determining the outcomes of interventions to prevent respiratory syncytial virus disease in children: what to measure? Lancet Respir Med. 2018;6(1):65–74.
15. Brown PM, Schneeberger DL, Piedimonte G. Biomarkers of respiratory syncytial virus (RSV) infection: specific neutrophil and cytokine levels provide increased accuracy in predicting disease severity. Paediatr Respir Rev. 2015;16(4):232–40.
16. Jong VL, Ahout IM, van den Ham H-J, Jans J, Zaaraoui-Boutahar F, Zomer A, Simonetti E, Bijl MA, Brand HK. van IJcken WF: Transcriptome assists prognosis of disease severity in respiratory syncytial virus infected infants. Sci Rep. 2016;6(1):1–12.
17. de Steenhuijsen Piters WA, Heinonen S, Hasrat R, Bunsow E, Smith B, Suarez-Arrabal MC, Chaussabel D, Cohen DM, Sanders EA, Ramilo O, *et al*. Nasopharyngeal microbiota, host transcriptome, and disease severity in children with respiratory syncytial virus infection. Am J Respir Crit Care Med. 2016;194(9):1104–15.
18. Mariani TJ, Qiu X, Chu C, Wang L, Thakar J, Holden-Wiltse J, Corbett A, Topham DJ, Falsey AR, Caserta MT, *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. 2017;216(8):1027–37.
19. Mejias A, Dimo B, Suarez NM, Garcia C, Suarez-Arrabal MC, Jartti T, Blankenship D, Jordan-Villegas A, Ardura MI, Xu Z, *et al*. Whole blood gene expression profiles to assess pathogenesis and disease severity in infants with respiratory syncytial virus infection. PLoS Med. 2013;10(11):e1001549.
20. Do LAH, Pellet J, van Doorn HR, Tran AT, Nguyen BH, Tran TTL, Tran QH, Vo QB, Tran Dac NA, Trinh HN, *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. 2017;217(1):134–46.
21. Walsh EE, Mariani TJ, Chu C, Grier A, Gill SR, Qiu X, Wang L, Holden-Wiltse J, Corbett A, Thakar J, *et al*. Aims, study design, and enrollment results from the assessing predictors of infant respiratory syncytial virus effects and severity study. JMIR Res Protoc. 2019;8(6):e12907.
22. Chu CY, Qiu X, Wang L, Bhattacharya S, Lofthus G, Corbett A, Holden-Wiltse J, Grier A, Tesini B, Gill SR, *et al*. The healthy infant nasal transcriptome: a benchmark study. Sci Rep. 2016;6:33994.
23. Caserta MT, Qiu X, Tesini B, Wang L, Murphy A, Corbett A, Topham DJ, Falsey AR, Holden-Wiltse J, Walsh EE. Development of a global respiratory severity score for respiratory syncytial virus infection in infants. J Infect Dis. 2017;215(5):750–6.
24. Cheng L, Lo LY, Tang NL, Wang D, Leung KS. CrossNorm: a novel normalization strategy for microarray data in cancers. Sci Rep. 2016;6:18898.
25. Cheng L, Wang X, Wong PK, Lee KY, Li L, Xu B, Wang D, Leung KS. ICN: a normalization method for gene expression data considering the overexpression of informative genes. Mol Biosyst. 2016;12(10):3057–66.
26. Liu X, Li N, Liu S, Wang J, Zhang N, Zheng X, Leung KS, Cheng L. Normalization methods for the analysis of unbalanced transcriptome data: a review. Front Bioeng Biotechnol. 2019;7:358.
27. Liu X, Zheng X, Wang J, Zhang N, Leung KS, Ye X, Cheng L. A long noncoding RNA signature for diagnostic prediction of sepsis upon ICU admission. Clin Transl Med. 2020, 10(3).
28. Zou H, Hastie T. Regularization and variable selection via the elastic net (vol B 67, pg 301, 2005). J R Stat Soc B. 2005;67:768–768.
29. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1–22.
30. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B. 1995;57:289–300.
31. Walsh EE, Wang L, Falsey AR, Qiu X, Corbett A, Holden-Wiltse J, Mariani TJ, Topham DJ, Caserta MT. Virus-specific antibody, viral load, and disease severity in respiratory syncytial virus infection. J Infect Dis. 2018;218(2):208–17.
32. Wright PF, Gruber WC, Peters M, Reed G, Zhu Y, Robinson F, Coleman-Dockery S, Graham BS. Illness severity, viral shedding, and antibody responses in infants hospitalized with bronchiolitis caused by respiratory syncytial virus. J Infect Dis. 2002;185(8):1011–8.
33. Yan XL, Li YN, Tang YJ, Xie ZP, Gao HC, Yang XM, Li YM, Liu LJ, Duan ZJ. Clinical characteristics and viral load of respiratory syncytial virus and human metapneumovirus in children hospitaled for acute lower respiratory tract infection. J Med Virol. 2017;89(4):589–97.
34. Piedra FA, Mei M, Avadhanula V, Mehta R, Aideyan L, Garofalo RP, Piedra PA. The interdependencies of viral load, the innate immune response, and clinical outcome in children presenting to the emergency department with respiratory syncytial virus-associated bronchiolitis. PLoS ONE. 2017;12(3):e0172953.
35. Collins PL, Fearns R, Graham BS. Respiratory syncytial virus: virology, reverse genetics, and pathogenesis of disease. Curr Top Microbiol Immunol. 2013;372:3–38.
36. Chu CY, Qiu X, McCall MN, Wang L, Corbett A, Holden-Wiltse J, Slaunwhite C, Grier A, Gill SR, Pryhuber GS et al. Airway gene expression correlates of RSV disease severity and microbiome composition in infants. J Infect Dis (in press).
37. Seber GA, Lee AJ. Linear regression analysis, vol. 329. Hoboken: Wiley; 2012.
38. Opsomer J, Wang Y, Yang Y. Nonparametric regression with correlated errors. Stat Sci. 2001:134–153
39. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. Statistics surveys. 2010;4:40–79.
40. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118–27.
41. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3(9):1724–35.
42. Rudy J, Valafar F. Empirical comparison of cross-platform normalization methods for gene expression data. BMC Bioinform. 2011;12:467.
43. Qiu X, Hu R, Wu Z. Evaluation of bias-variance trade-off for commonly used post-summarizing normalization procedures in large-scale gene expression studies. PLoS ONE. 2014;9(6):e99380.
44. Lee SH, Ruan SY, Pan SC, Lee TF, Chien JY, Hsueh PR. Performance of a multiplex PCR pneumonia panel for the identification of respiratory pathogens and the main determinants of resistance from the lower respiratory tract specimens of adult patients in intensive care units. J Microbiol Immunol Infect. 2019;52(6):920–8.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.