Integrated, Multi-cohort Analysis Identifies Conserved Transcriptional Signatures across Multiple Respiratory Viruses

Marta Andres-Terre, Helen M. McGuire, Yannick Pouliot, Erika Bongen, Timothy E. Sweeney, Cristina M. Tato, and Purvesh Khatri

METHODS

Data Collection and Pre-Processing
Our entire analysis was performed using publicly available data. We downloaded 15 microarray gene expression datasets from the NCBI GEO comprising of 2,785 samples derived from whole blood, PBMCs, epithelial cells or cell lines (Supplementary Table 1). The samples in these datasets represented different biological conditions including viral infections (influenza, RSV, HRV, SARS, adenovirus, enterovirus, HHV6), bacterial infections (*E. coli, S. aureus, S. pneumonie, Salmonella*), non-pathogenic systematic inflammatory response, and healthy controls. We incorporated technical heterogeneity in our analysis by choosing datasets that were profiled using microarrays from different manufacturers. All datasets, except one (GSE19392), are whole blood or PBMC samples obtained from patients with or without a viral infection over wide range of age (from less than 2 months to over 60 years). Furthermore, the samples were independently collected and profiled at 14 centers in seven countries.

For all datasets, we verified that the expression data available for download from the NCBI GEO database was normalized and log2-transformed. We identified that in GSE17156 samples for a given viral infection (influenza, RSV or HRV) were normalized independently of the other viral infections in the dataset. Therefore, we downloaded the raw data files for this GSE from GEO and normalized all samples within the dataset together using gcRMA. For each study, we used the sample phenotypes as defined by the primary publication of a source study. Microarray probes in each data set were mapped to Entrez Gene identifiers (IDs) to facilitate integrated analysis. If a probe matched more than one gene, the expression data for that probe were expanded to add one record for each mapped gene[9].

Dataset selection for MVS and IMS analysis
We had two major hypotheses in this paper: first, that there exists a common host response to clinical respiratory viral infection that is not specific to virus type and may be reflective of illness severity; and second, that virus-specific signatures also exist and can separate different viral infections. To that end, we performed two multi-cohort analyses, which we termed 'meta-virus signature' (MVS) and 'influenza meta-signature' (IMS). The MVS analysis incorporated datasets from multiple respiratory viruses, while the IMS analysis examined only influenza.

Discovery cohort selection for MVS analysis
Unlike a conventional biological experiment, we sought to represent biological and technical heterogeneity observed in the population by choosing cohorts that were collected at different centers and profiled using different technologies. We chose three datasets (GSE34205, GSE42026, and GSE40396), from which we created five discovery cohorts. Each cohort contained samples from only one type of viral infection and healthy controls. For instance, we created two cohorts from GSE34205: (1) comparing 22 healthy control samples with 28 samples from patients with influenza infection, and (2) comparing 22 healthy control samples with 51

samples from patients with RSV infection. The 22 healthy control samples were common between these two cohorts. Similarly, we created two cohorts from GSE42026 that compared samples from influenza and RSV infected patients with healthy controls. Finally, we compared only HRV samples with healthy controls in GSE40396. The remaining samples with viral or bacterial infections were used as a validation cohort for the MVS in Figure 2A.

Discovery cohort selection for IMS analysis
Similar to the approach we took for selecting data sets in discovering the MVS signature, we again sought to represent biological and technical heterogeneity observed in the population. We represented the phenotypic diversity by choosing samples from the following cohorts as "controls": healthy, day 0 (pre-inoculation) and bacterial pneumonia; and "cases": post-inoculation and influenza pneumonia. We chose datasets that were collected at centers in different countries to represent different biological confounding factors such as population genetics, treatment protocols and infectious agents. We also incorporated technical heterogeneity in our samples as before, by choosing datasets that were profiled using different microarray platforms.

We identified five datasets consisting of 292 blood samples that satisfied these criteria. In GSE17156, all pre-inoculation samples with influenza were used as control samples, and post-inoculation samples with influenza were used as cases, irrespective of whether they were symptomatic or asymptomatic after inoculation. In GSE20436, all bacterial pneumonia samples and pre-vaccination samples were used as controls, and severe influenza or day 7 post-vaccination samples were used as cases. In GSE40012, only samples from patients with influenza pneumonia and healthy controls were used, irrespective of which day the samples were taken; the rest of the samples in this dataset (bacterial pneumonia, mixed-pneumonia, and non-infectious systematic inflammatory response) were used for validation of the IMS in Figure 5A and 7F. In GSE34205 and GSE42026, we used samples from healthy controls and influenza-infected samples, but samples from RSV-infected patients were used for validating specificity of the IMS to influenza infection in Figure 6A and 6B.

Meta-analysis by Combining Effect Size
We applied two meta-analysis methods as described in our previous publication (Supplementary Figure 1; Supplementary Table 1)[8]. In this framework, for each analysis a subset of the total 15 datasets was selected to be used as discovery cohorts as described above (Supplementary Table 2), and analyzed using two meta-analysis methods: i) combining effect sizes and ii) combining P-values. We estimated the effect size for each gene in each data set as Hedges' adjusted *g*. If multiple probes mapped to a gene, the effect size for each gene was summarized using the fixed effect inverse-variance model. The study-specific effect sizes for each gene were then combined into a single meta effect-size using a linear combination of study-specific effect sizes, f_i, where each study-specific effect size was weighted by inverse of the variance in the corresponding study. After computing meta effect-size, p-values were

corrected for multiple hypotheses testing using Benjamini-Hochberg false discovery rate (FDR) correction[10], and significant genes were identified using Z-statistic.

Meta-analysis by Combining P-values

We used Fisher's sum of logs method[11] for meta-analysis by combining p-values. For each gene, we summed the logarithm of the one-sided hypothesis testing p-values across k studies, and compared the result to a χ^2-distribution with $2k$ degrees of freedom.

Leave-One-Out Analysis

A dataset with a large sample size can significantly influence meta-analysis results. Therefore, we performed meta-analysis by removing one dataset at a time in order to avoid influence of a single dataset in identifying the MVS and IMS. We applied both meta-analysis methods at each iteration. Using FDR ≤ 1% and FDR ≤ 0.01% at each iteration for MVS and IMS, respectively, we identified 396 differentially expressed genes (161 over-expressed, 235 under-expressed) across all viral infections, and 11 over-expressed genes across all influenza-infection datasets. We allowed significant heterogeneity when identifying MVS as different viruses may induce various genes at different levels. However, we ensured that there was no significant heterogeneity (P > 0.05) in effect size across all datasets for the IMS.

Meta Virus Signature and Influenza Meta-Signature Score

We defined the MVS score of a sample as the geometric mean of the normalized, log2-transformed expression of the 161 over-expressed genes minus that of the 235 under-expressed genes. We defined the IMS score of a sample as the geometric mean of the normalized log2-transformed expression of the 11 over-expressed genes. We scaled and centered the MVS or IMS scores of all samples in a given dataset (mean = 0, standard deviation = 1) to enable comparisons between datasets. We used the Wilcoxon–Mann–Whitney[12] or ANOVA to test whether there was a statistically significant difference between the MVS or IMS scores of two groups.

If a dataset contained negative values, computing a geometric mean is not possible. In these datasets, we used mean of the normalized log2-transformed expression values to compute the MVS and IMS scores.

Pathway Analysis

We performed functional pathway analysis using iPathwayGuide[13-15]. Meta-effect size across all viral infections was used as fold change in iPathwayGuide to identify significant pathways. We used FDR ≤ 10% as a threshold for identifying significant pathways. We performed network analysis of influenza-specific genes using IPA with an option to include only "direct relationship" to avoid spurious connections caused by "indirect relations." Direct relationships in IPA result from publications citing experimental evidence for an interaction.

Study Summaries

This section describes each of the dataset used in the analysis. In order to accurately describe these datasets, their description has been used verbatim from their corresponding manuscript as much as possible.

GSE34205: Ioannidis *et al.*[16] profiled blood samples from children with a median range of 2.4 (range: 1.5-8.6) months that were hospitalized with acute RSV or influenza infection. Samples from these patients were collected within 42 to 72 hours of hospitalization. They excluded any children with suspected or proven polymicrobial infections, with underlying chronic medical conditions, with immunodeficiency, or those who received systemic steroids or other immunomodulatory therapies. Control samples were collected from children undergoing elective surgical procedures or at outpatient clinical visits. To exclude viral coinfections, they performed nasopharyngeal viral cultures of all subjects.

GSE42026: Herberg *et al.*[17] recruited febrile children (<17 years) with respiratory infection, and collected whole blood samples from them.

GSE6269: Ramilo *et al.*[4] profiled PBMCs from young patients (age ≤ 18 years) with acute infections caused by common human pathogens: (1) an RNA virus (influenza A); (2) two gram-positive bacteria (*Staphylococcus aureus* and *Streptococcus pneumoniae*); and (4) a Gram-negative bacterium (*Escherichia coli)*. The patients in this dataset were treated according to standard hospital protocols, and as such, antimicrobial therapy was promptly initiated in the emergency department. The patients were treated with up to three different drugs from an overall set of 13 drugs, which represented a potential confounding factor in our analysis. The study profiled these samples using three types of microarrays from two manufacturers, Affymetrix (HG U133A and HG U133 plus 2.0) and Illumina, representing technical variability.

GSE40012: Parnell *et al.*[2] recruited patients with severe community-acquired pneumonia requiring severe intensive care unit (ICU) admission. They also recruited patients with noninfective systemic inflammatory response syndrome (SIRS) and healthy controls. SIRS was defined as the presence of at least two of the following four clinical criteria: (a) fever or hypothermia (temperature > 100.4°F (38°C) or < 96.8°F (36°C)); (b) tachycardia (> 90 beats/min), (c) tachypnea (> 20 breaths/min or $PaCO2 < 4.3$ kPa (32 mm Hg)), or the need for mechanical ventilation; (d) an altered white blood cell count of > 12,000 cells/μl, < 4,000 cells/μl, or the presence of > 10% band forms. Pneumonia was defined as a microbiologically confirmed infection of the lungs, resulting in the patient fulfilling the SIRS criteria. The first sample from each patient was collected within the initial 24 hours of admission to the ICU, referred to as Day 1. Patients were monitored for upto 5 days to access their longitudinal gene-expression profiles.

GSE21802: Bermejo-Martin *et al.*[18] recruited patients with primary viral pneumonia during the acute phase of influenza virus illness with acute respiratory distress and unequivocal alveolar opacification involving two or more lobes with negative respiratory and blood bacterial cultures at admission to ICUs. They excluded patients older than 65 years and younger than 18 years from the study to avoid immaturity/aging of the immune system as confusion factor in the analysis. Only those patients with con- firmed H1N1 infection by real-time polymerase chain reaction (PCR) were included in the study. Healthy volunteers of similar age to the patients were recruited between workers of the University of Valladolid, Spain. Treatment decisions for all patients, including corticosteroid therapy, were not standardized and were decided by the attending physician.

GSE20346: Parnell *et al.*[3] recruited critically ill patients with severe infection, defined as infection where there is at least one major organ failure that requires critical care intervention. These patients had either viral infections (seasonal H3N2 or pandemic H1N1/09 influenza virus) or bacterial infections. These patients were followed for a further four days. Healthy volunteers were enrolled from a local influenza vaccination program.

GSE17156: Zaas *et al.*[1] inoculated healthy volunteers with one of the three viruses (HRV, RSV, influenza) as described below.

HRV cohort: Zaas *et al.* recruited healthy volunteers through an active screening protocol at the University of Virginia (Charlottesville, VA). On the day of inoculation, 10^6 $TCID_{50}$ GMP HRV serotype 39 was inoculated intranasally. Subjects were admitted to the quarantine facility for 48 hours following HRV inoculation and remained for 48 hours following inoculation. Nasopharyngeal (NP) lavage samples were obtained from each subject daily for HRV titers to accurately gauge the success and timing of the HRV inoculation. Following the 48th hour after inoculation, subjects were released from quarantine and returned for 3 consecutive mornings for sample acquisition.

RSV cohort: The RSV challenge was performed by Retroscreen Virology, Ltd. (London). On the day of inoculation, a dose of 10^4 $TCID_{50}$ RSV (serotype A) was inoculated intranasally per standard methods. Blood and NP lavage collection methods were similar to the HRV cohort, but continued throughout the quarantine. Due to the incubation period of RSV A, subjects were not released from quarantine until after the 288th hour and were negative by rapid RSV antigen detection.

Influenza cohort: A healthy volunteer intranasal challenge with influenza A/Wisconsin/67/2005 (H3N2) was performed at Retroscreen Virology, Ltd. (Brentwood, UK). On the day of inoculation, a dose of 10^6 $TCID_{50}$ influenza A was diluted and inoculated intranasally per standard methods at a varying dose (1:10, 1:100, 1:1000, 1:10000) with four or five subjects receiving each dose. Due to the incubation period, subjects were not released from

quarantine until after the 168th hour. Blood and NP lavage collection continued throughout the duration of the quar- antine. All subjects received oral oseltamivir (Roche Pharmaceuticals) (75 mg) by mouth twice daily at day 6 following inoculation and were negative by rapid antigen detection at time of discharge.

GSE38900: Mejias *et al.*[19] performed a prospective observational study over six respiratory seasons including a cohort of hospitalized infants (<2 years old) with RSV, HRV, and influenza infections in one of three centers: (1) Nationwide Children's hospital, Columbus, Ohio, (2) Turku University (Turku, Finland), and (3) Children's Medical Center, Dallas, Texas, US. Children with documented bacterial co-infections (including bacteremia, urinary tract infection, meningitis, acute gastroenteritis, or any bacterial pathogen isolated from a sterile site) or viral co-infections were excluded from the study. They also excluded children with congenital heart disease, chronic lung disease, immunodeficiency, prematurity (<36 weeks), and systemic steroid treatment within 2 weeks before presentation. Control samples were obtained from healthy children undergoing elective surgery not involving the respiratory tract, or at routine outpatient visits. For the healthy control group, a clinical questionnaire was used, and those children with co-morbidities, use of systemic steroids, or presence of any illness within 2 weeks prior to enrollment were excluded. Lastly, to exclude viral co-infections, respiratory samples were tested by viral culture or PCR in 94% of patients and controls.

GSE40396: Hu *et al.*[20] profiled blood samples from 65 children under three years of age with bacterial (n=8) or viral (n=35) infections, and 22 virus-negative afebrile controls. All children with bacterial infections were febrile, whereas only a subset of children (22/35) with viral infections were febrile. For instance, all children infected with rhinovirus were afebrile.

GSE1739: Reghunathan *et al.*[21] recruited adult patients who were diagnosed with severe acute respiratory syndrome (SARS) according to the World Health Organization (WHO) SARS criteria and admitted to the Tan Tock Seng Hospital, Singapore.

GSE52428: Woods *et al.*[22] inoculated adult volunteers (20-41 years) with either H1N1 (A/Brisbane/59/2007) or H3N2 (A/Wisconsin/67/2005) influenza strains, and collected whole blood samples at 8 hours intervals after inoculation. Symptoms were recorded twice daily using a modified standardized symptom score, called the modified Jackson Score, which requires subjects to rank symptoms of upper respiratory infection (stuffy nose, scratchy throat, headache, cough, etc) on a scale of 0– 3 of ''no symptoms'', ''just noticeable'', ''bothersome but can still do activities'' and ''bothersome and cannot do daily activities''. For all cohorts, modified Jackson scores were tabulated to determine if subjects became symptomatic from the respiratory viral challenge. Symptom onset was defined as the first of 2 contiguous days with score of 2 or more. A modified Jackson score of ≥6 over a consecutive five day period was the primary indicator of symptomatic viral infection and subjects with this score and a positive qualitative viral culture or quantitative RT-PCR for at least 2 consecutive days (beginning 24

hours after inoculation) were denoted as ''symptomatic infection''. Subjects were classified as ''asymptomatic, not infected'' if the symptom score was less than 6 over the five days of observation and viral shedding was not documented after the first 24 hours subsequent to inoculation as above. Standardized symptom scores were tabulated at the end of each study to determine attack rate and time of maximal symptoms.

GSE47353: Tsang et al.[23] Healthy volunteers over the age of 18 were enrolled on the National Institutes of Health (NIH) protocol 09-H-0239 (Clinicaltrials.gov NCT01191853). Healthy volunteers were screened for protocol enrollment with a medical history, physical examination, and clinical laboratory studies (CBC with differential, blood chemistry, coagulation and thrombosis screens (PT, PTT, D-dimer), cholesterol panel, urinalysis, and pregnancy test). Pregnant individuals and those who had received vaccines or taken immune modifying medications within six months of study entry were excluded. Individuals meeting inclusion criteria were vaccinated with the 2009 Fluvirin seasonal influenza (Novartis), and H1N1 pandemic (Sanofi-Aventis) vaccines, both without adjuvant. Blood samples were obtained at day -7 (50ml) prior to vaccination, day 0 (150ml) immediately before vaccination, day1 (40ml), day7 (150ml), and day70 (150ml) post vaccination. All blood samples were drawn between 8am and 11am from fasting individuals and were processed within 30 minutes of drawing.

GSE48018: Franco et al.[24] enrolled healthy volunteers ages 18 to 40 years. Individuals who were known to have received an influenza vaccine in the previous 3 years or who had signs or symptoms of an active infection at the time of enrollment were excluded. To minimize false-positive results related to population stratification, enrollment was limited to individuals of self-reported Caucasian ancestry. All subjects in this cohort were males. Study participants were immunized on day 0. Those enrolled in the initial cohort received the 2008–2009 inactivated trivalent influenza vaccine (A/Brisbane/59/2007[H1N1], A/Brisbane/10/2007[H3N2], B/Florida/4/2006; Sanofi-Pasteur, Lyon, France).

GSE48023: This cohort was generated as a validation cohort of GSE48018 by Franco et al.[24] following the same protocol as GSE48018. However, unlike GSE48018, all subjects in this cohort were females, and was collected 18 months after GSE48018. Individuals in this cohort received the 2009–2010 vaccine, which came from the same manufacturer as GSE48018 and included (A/Brisbane/59/2007), (A/Brisbane/ 10/2007[H3N2]), and (B/Brisbane/60/2008) strains.
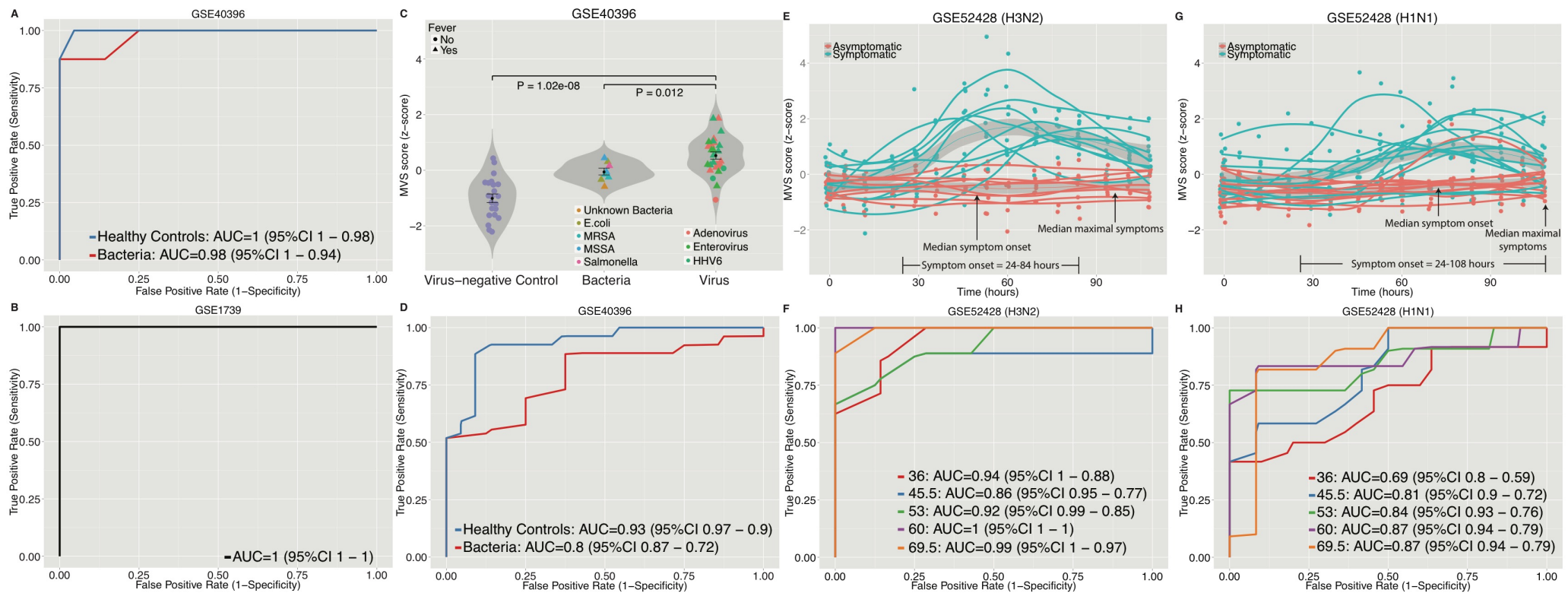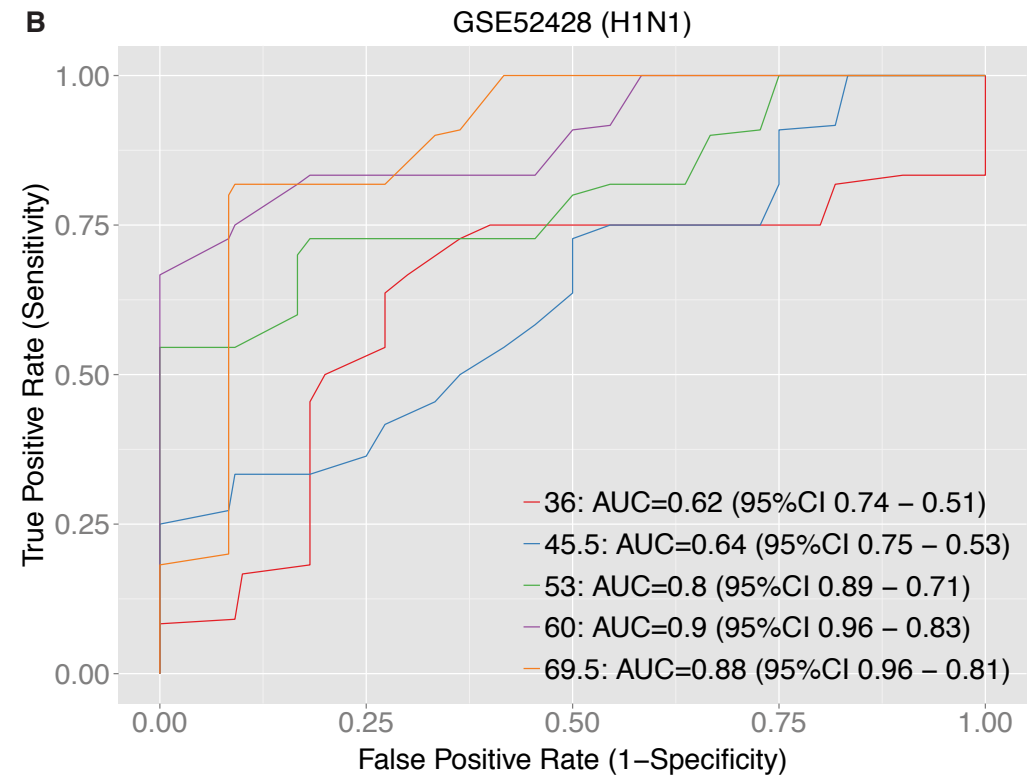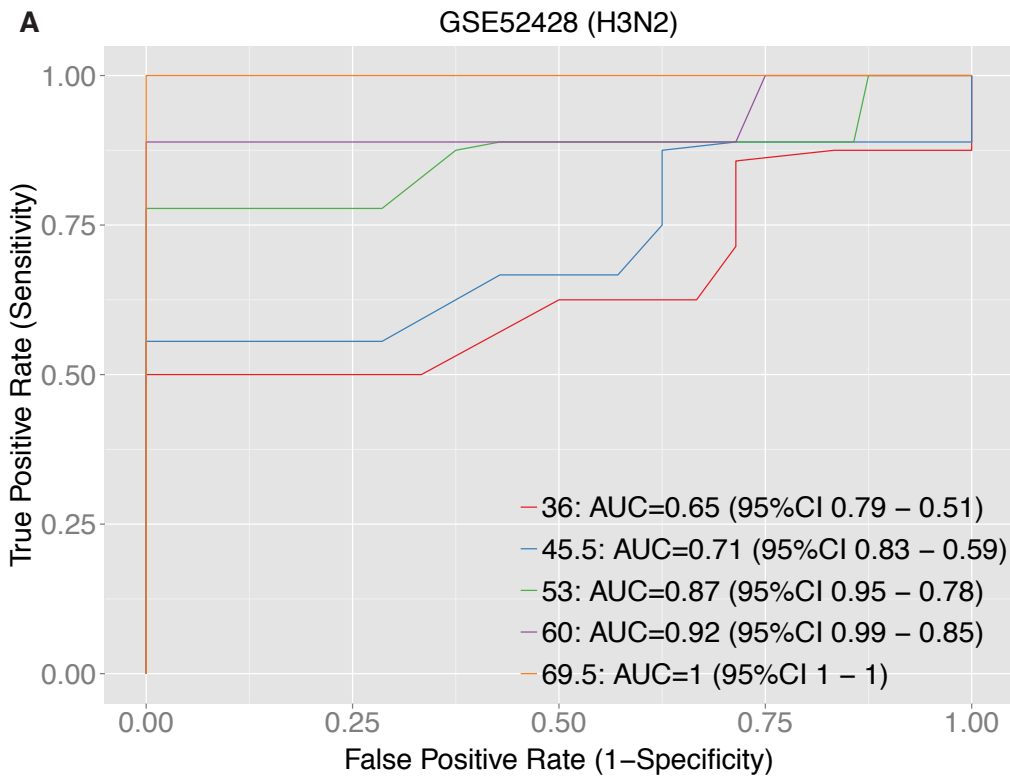
Supplementary Fig 1, refers to Fig 2 : Performance of the MVS in different viral infections. (**a**) ROCs of MVS scores distinguishing patients with virus infection from those with bacterial infection and healthy controls. (**b**) ROC of MVS score distinguishing patients with SARS coronavirus infection from healthy controls. (**c-d**) MVS scores in symptomatic and asymptomatic subjects inoculated with H3N2 or H1N1. Each line represents MVS score for an individual. Smoothed curves for each individual are drawn using LOESS. Gray bars indicate 95% confidence interval. (**e-f**) ROCs of MVS scores for distinguishing symptomatic and asymptomatic individuals at different time points after influenza A inoculation. Panel "d" is missing in this figure.
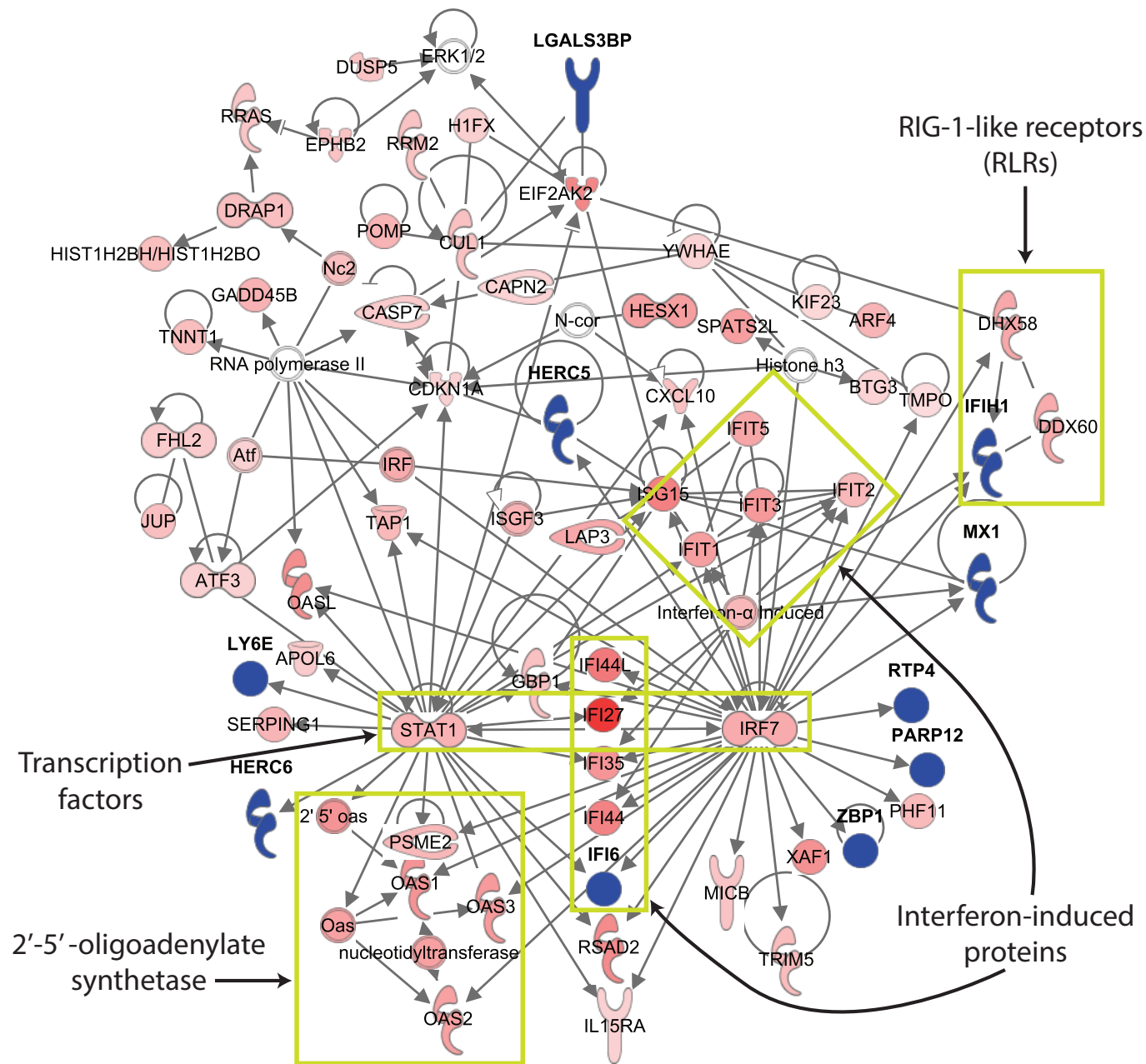
Supplementary Fig 2, refers to Fig 3: Network analysis of the 127-gene Influenza signature using Ingenuity Pathway Analysis. Of the 127 identified genes, 71 are involved in innate virus sensing and initiation of antiviral response pathways. Yellow boxes indicate transcription factors (STAT1, IRF7), RIG-1-like receptors (RLRs), 2'-5'-oligoadenylate synthetases and Interferon-induced proteins that have been previously implicated in response to influenza infection.
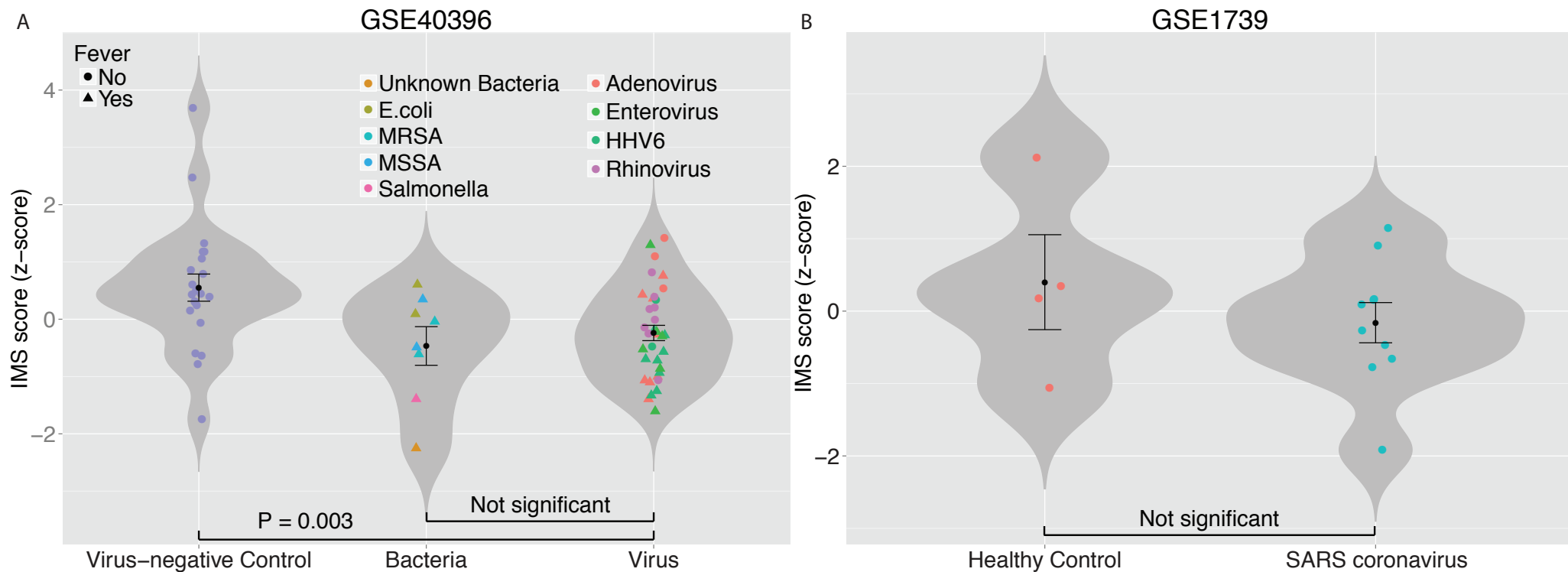
Supplementary Fig 3, refers to Fig 4 and 5: Comparison of IMS scores in healthy controls and patients with bacterial infection to those with viral infection other than influenza. (**a**) IMS scores in febrile and afebrile pediatric patients with viral or bacterial infections. (**b**) IMS scores in SARS-infected patients and healthy controls. Error bars indicate mean ± standard error for a given group of samples. Width of a violin plot indicates density of samples, where each dot represents a sample.

Supplementary Fig 4, refers to Fig 6: Performance of IMS in distinguishing symptomatic and asymptomatic subjects. ROCs of IMS scores in distinguishing symptomatic and asymptomatic subjects inoculated with (**a**) Influenza H3N2 or (**b**) Influenza H1N1.
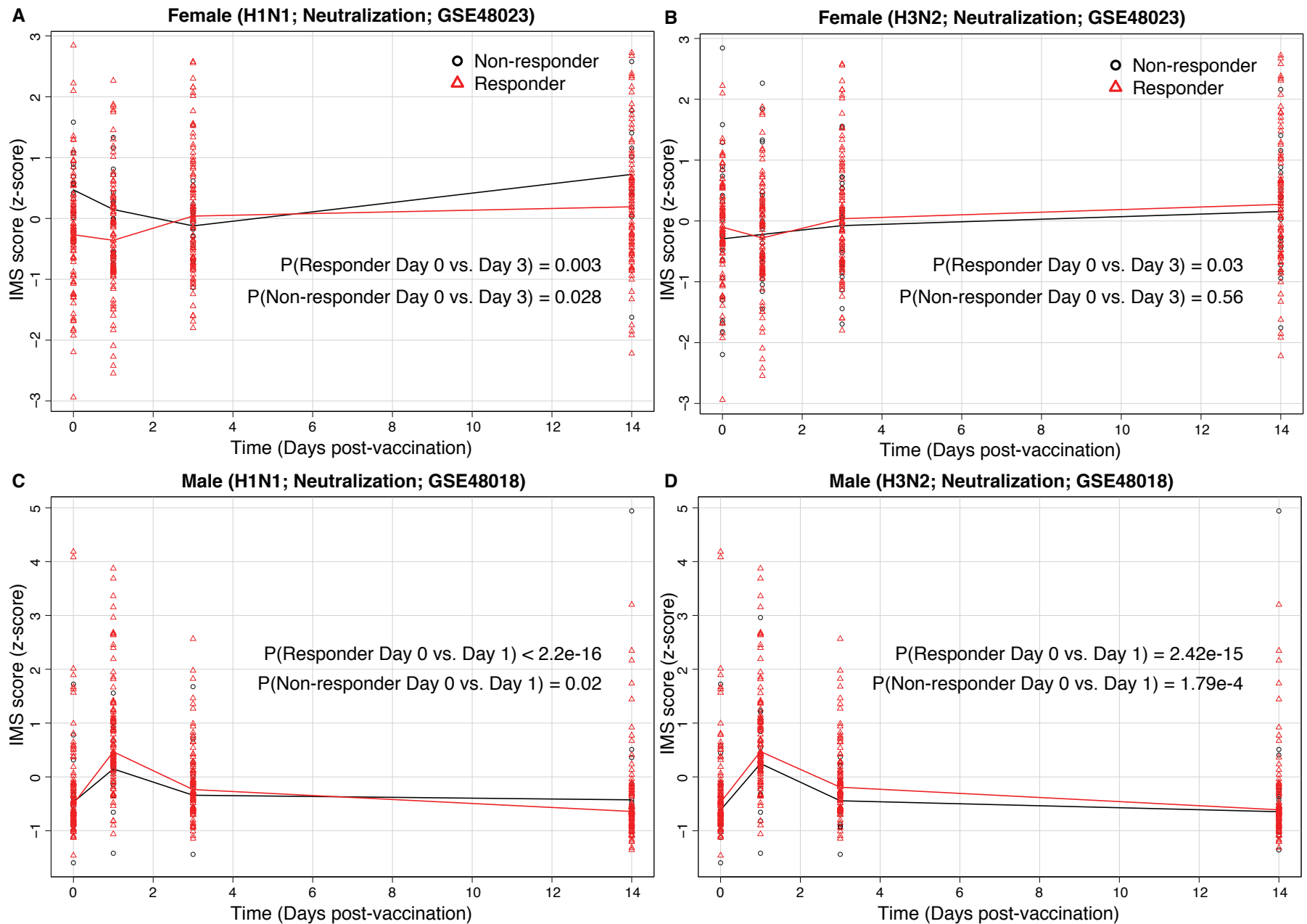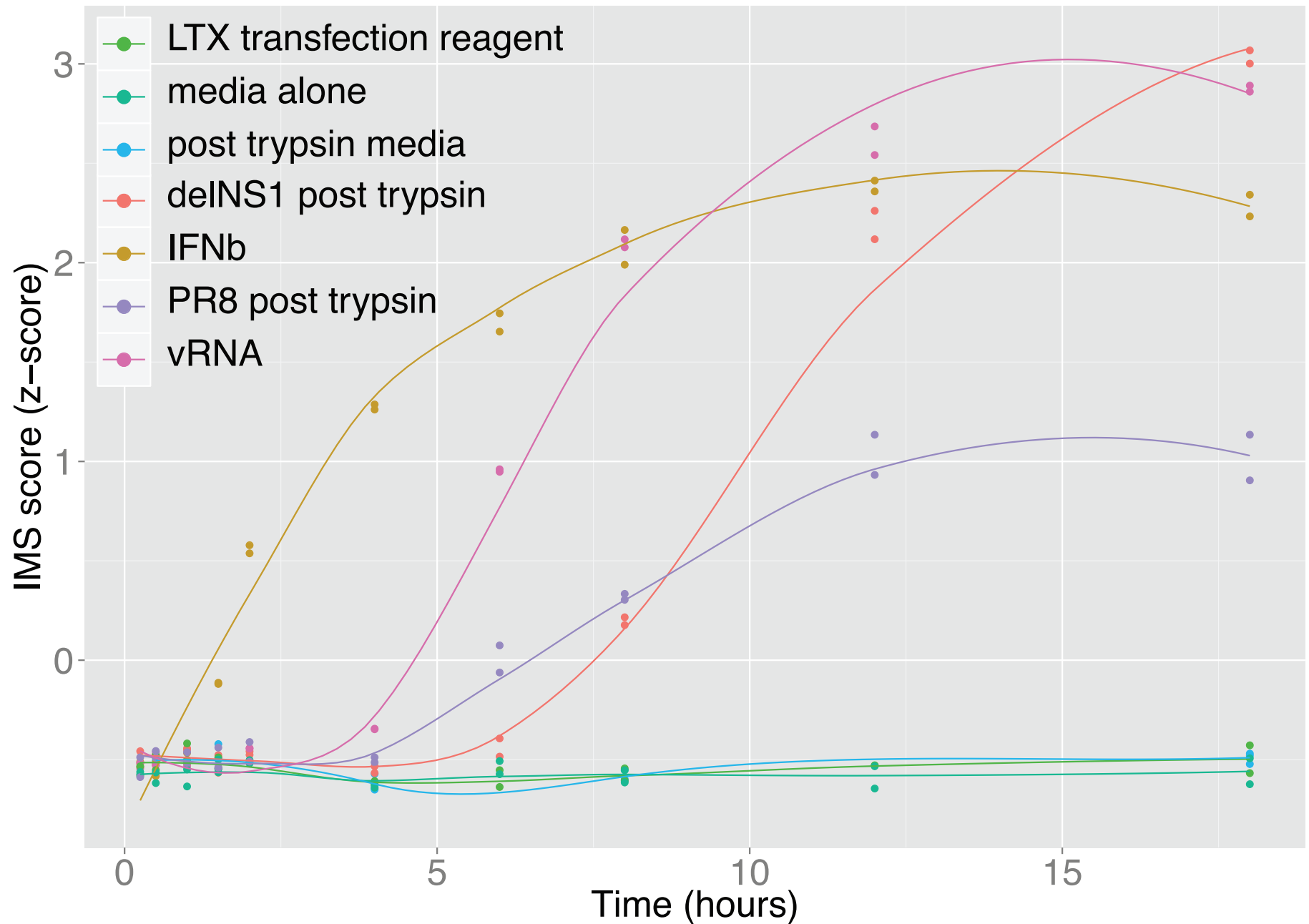
Supplementary Fig 5, refers to Fig 7: IMS score increases significantly in vaccine responders when defined using microneutralization titers instead of HAI titers. (*a-b*) Change in IMS scores for vaccine responders and non-responders, defined based on microneutralization titers, in a female cohort following influenza vaccination. (*c-d*) Change in IMS scores for vaccine responders and non-responders, defined based on microneutralization titers, in a male cohort following influenza vaccination.
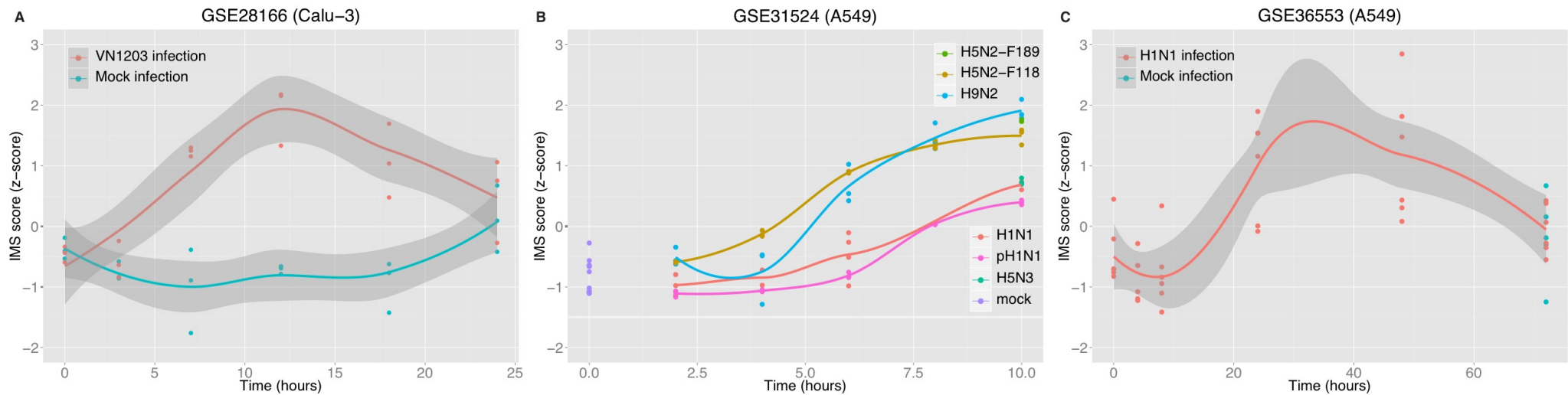
Supplementary Fig 6, refers to Fig 3: IMS scores in epithelial cell lines infected with different influenza virus strains. (**a**) Calu-3 cells were infected with Influenza A/VN/1203. (**b**) A549 cells were infected with Influenza H5N2-F189, H5N2-F118, H9N2, H1N1, pH1N1, H5N3 or mock-control and monitored until 10h post-infection. (**c**) A549 cells were infected with Influenza H1N1 or mock infected and monitored until 70h post-infection. Smoothed lines indicate loess curves for infected or mock-infected cells. Gray bars indicate 95% confidence interval.

Supplementary Fig 7, refers to Fig 6: Change in IMS scores in primary human bronchial epithelial cells following infection with wild type influenza or NS1-deleted mutant influenza, transfection with viral RNA, and treatment with IFNβ.