Supplemental Data
Cell Host & Microbe, *Volume 6*

Gene Expression Signatures Diagnose Influenza and Other Symptomatic Respiratory Viral Infections in Humans

Aimee K. Zaas, Minhua Chen, Jay Varkey, Timothy Veldman, Alfred O. Hero III, Joseph Lucas, Yongsheng Huang, Ronald Turner, Anthony Gilbert, Robert Lambkin-Williams, N. Christine Øien, Bradly Nicholson, Stephen Kingsmore, Lawrence Carin, Christopher W. Woods, and Geoffrey S. Ginsburg

Supplemental Experimental Procedures

Institutional Review Board Approvals

Rhinovirus: The protocol was approved by the Human Investigations Committee of the University of Virginia, the IRB of Duke University Medical Center, and the SSC-SD IRB (US Department of Defense; Washington, DC).

RSV: The protocol was approved by the East London and City Research Ethics Committee 1 (London, UK), an independent institutional review board (WIRB: Western Institutional Review Board; Olympia, WA), the IRB of Duke University Medical Center (Durham, NC), and the SSC-SD IRB (US Department of Defense; Washington, DC).

Influenza: The protocol was approved by the East London and City Research Ethics Committee 1 (London, UK), an independent institutional review board (WIRB: Western Institutional Review Board; Olympia, WA), the IRB of Duke University Medical Center (Durham, NC), and the SSC-SD IRB (US Department of Defense; Washington, DC).

RNA Purification

Briefly, the samples were removed from –80°C and incubated at room temperature for 2 hr to ensure complete lysis. Following lysis, the tubes were centrifuged for 10 min at 5,000 × g, the supernatant was decanted, and 4 ml of RNase-free water was added to the pellet. The tube was vortexed to thoroughly resuspend the pellet and centrifuged for 10 min at 5000 × g, and the entire supernatant was discarded. The remaining pelleted lysate was resuspended in 350 µl of buffer BR1 by vortexing, added to microcentrifuge tubes containing 40 µl of Proteinase K and 300 µl of buffer BR2, then incubated for 30 min at 65°C. Following incubation, the specimens were transferred to a PAXgene 96 Filter Plate and centrifuged at 5600 x g for 10 min. Three hundred fifty microliters of 100% ethanol was added to the eluate from the filter plate and mixed by gentile pipetting. The entire volume was then transferred to a PAXgene 96 RNA plate with negative pressure applied via vacuum manifold. The plate was then washed by adding 500 µl of buffer BR3 per well under negative pressure. Eighty microliters DNase I incubation mix was added directly to the PAXgene membrane, then allowed to incubate at room temperature and pressure. Following DNase I digestion, the plate was washed with successive applications of 500 µl buffer BR3 and 1 ml BR4 with negative pressure. A final volume of 1 ml of buffer BR4 was added to each membrane, then the plate was centrifuged at

5600 x g for 10 min to ensure the membranes were completely dry. RNA was eluted in two applications of 45 μL buffer BR5 via centrifugation at 5600 x g for 4 min for each elution. RNA was quantified via UV spectrophotometer, and quality confirmed via the Agilent 2100 Bioanalyzer (Agilent; Santa Clara, CA).

These arrays contain probes for approximately 18,400 transcripts and variants, including over 14,500 well-characterized human genes. The sequences from which these probe sets were derived were selected from GenBank, dbEST, and RefSeq. The sequence clusters were created from the UniGene database (Build 133, April 20, 2001) and then refined by analysis and comparison with a number of other publicly available databases, including the Washington University EST trace repository and the University of California, Santa Cruz Golden-Path human genome database (April 2001 release). Some gene sequences were selected from GenBank, dbEST, and RefSeq. Sequence clusters were created from the UniGene database (Build 159, January 25, 2003) and refined by analysis and comparison with a number of other publicly available databases, including the Washington University EST trace repository and the NCBI human genome assembly (Build 31).

Target was prepared and hybridized according to the Affymetrix Technical Manual. Total RNA (2.5 μg) was converted into cDNA using Reverse Transcriptase (Invitrogen) and a modified oligo(dT)24 primer that contains T7 promoter sequences (GenSet). A set of four peptide nucleic acid (PNA) oligomers (Applied Biosystems; Foster City, CA) with sequences complimentary to globin mRNA were added to reduce globin RNA transcription. After first-strand synthesis, residual RNA was degraded by the addition of RNaseH, and a double-stranded cDNA molecule was generated using DNA Polymerase I and DNA Ligase. The cDNA was then purified and concentrated using a phenol:chloroform extraction followed by ethanol precipitation. The cDNA products were incubated with T7 RNA Polymerase and biotinylated ribonucleotides using an In Vitro Transcription kit (Affymetrix). The resultant cRNA product was purified using an RNeasy column (QIAGEN) and quantified with a spectrophotometer. The cRNA target (20 μg) was incubated at 94ºC for 35 min in fragmentation buffer (Tris, MgOAc, KOAc). The fragmented cRNA was diluted in hybridization buffer (MES, NaCl, EDTA, Tween 20, Herring Sperm DNA, Acetylated BSA) containing biotin-labeled OligoB2 and Eukaryotic Hybridization Controls (Affymetrix). The hybridization cocktail was denatured at 99°C for 5 min, incubated at 45°C for 5 min, and then injected into a GeneChip cartridge. The GeneChip array was incubated at 42°C for at least 16 hr in a rotating oven at 60 rpm. GeneChips were washed with a series of nonstringent (25°C) and stringent (50°C) solutions containing variable amounts of MES, Tween20 and SSPE. The microarrays were then stained with Streptavidin Phycoerythrin and the fluorescent signal was amplified using a biotinylated antibody solution. Fluorescent images were detected in an GeneChip Scanner 3000 and expression data was extracted using the GeneChip Operating System v 1.1 (Affymetrix). All GeneChips were scaled to a median intensity setting of 500.

Supplemental Analysis

Using just the flu data, we tested (Kruskal-Wallis) each probe for differential expression between subjects who were sick versus healthy at t_{max}. We note that, due to the small sample size, there were no probes showing significant association after correction for multiple hypotheses (Bonferroni). However, the quantile-quantile plot of the associated p-values (Figure S8) demonstrates significant deviation below the diagonal, indicating that there is a large predominance of low p-values. Clustering of all probe sets with p-value less than 0.001 (Figure S9) groups all sick patients in a cluster separate from healthy, independent of the time of measurement, and shows significant coexpression of many genes across the samples. Similar analyses of the data produced in the rhinovirus and RSV studies produced essentially identical results.

We analyzed jointly the results from all three trials in an ANOVA framework. In addition to the intercept term, we included in the design matrix indicators of sick versus healthy, t_0 versus t_{max}, and indicator for each of rhinovirus and RSV, and interaction terms for rhino-sick and RSV-sick. Among the 22,215 genes on the array were 599 that showed significant association with the indicator of sick versus healthy (after correction for multiple hypotheses). In addition, there were 16 genes showing association with the t_{max} indicator, 4 with the rhino-sick interaction term, and 9 genes associated with the RSV-sick interaction term. Clustering on all of the genes with p-values less than 0.05/22,215 for any of these categories (of which there were 614) produces clusters that distinguish sick from healthy patients (Figure S10). There is no evidence of a distinction between healthy patients measured at different time points. The sick patients are grouped into two separate clusters: one of size 15, which contains all 9 of the flu samples, and the other of size 9, which contains 1 healthy patient. There are 3 subjects sick with Rhinovirus and 1 subject sick with RSV that are clustered with the healthy patients. The grouping of all sick flu subjects into the same cluster indicates the possibility of building a signature based on the host cell response that will distinguish flu from other viral infections. However, because these three studies were performed at different times, it is also possible that this distinction is due to batch effects from processing the expression arrays.

Table S1. Gene Lists in Rank Order as Contribution to Factor Loading

Rhinovirus[a]	**RSV**[a]	**Influenza**[a]	**Panviral**[b]	**Literature**[c]
RSAD2	RSAD2	RSAD2	RSAD2	SIGLEC1
LAMP3	SERPING1	IFI44L	IFI44L	IFI27
IFI44L	IFI44L	SIGLEC1	LAMP3	IFI44L
IFIT1	IFIT1	LAMP3	SERPING1	IFIT1
SIGLEC1	IFI44	IFIT1	IFI44	RSAD2
IFI44	IFI44	IFI44	IFIT1	LOC727996
OAS3	LAMP3	SERPING1	IFI44	IFIT3
LOC727996	OAS3	IFI27	ISG15	IFI44L
SERPING1	HERC5	ISG15	SIGLEC1	OTOF
HERC5	ISG15	IFI44	OAS3	ISG15
ISG15	IFIT3	HERC5	HERC5	MX1
IFI6	SIGLEC1	LOC26010	LOC727996	SIGLEC1
IFI44	OASL	IFI6	IFIT3	HESX1
INDO	OAS1	LOC727996	IFI6	SERPING1
MX1	LOC26010	IFIT3	OASL	DHX58
IFIT3	MX1	OAS3	IFI27	LAMP3
OASL	IFI6	OASL	ATF3	LY6E
LOC26010	FCGR1A	SEPT4	MX1	OAS3
OASL	GBP1	XAF1	OAS1	IFI44
CXCL10	ATF3	OAS1	LOC26010	HERC5
ATF3	IFIT5	LY6E	XAF1	OAS1
OAS1	LAP3	MS4A4A	IFIT2	HERC6
DDX58	OASL	SIGLEC1	OAS2	XAF1
OAS1	IFIT2	TNFAIP6	LY6E	IFI6
LY6E	RTP4	CCL2	SEPT4	OAS2
OAS2	GBP1	OAS1	DDX58	STAT1
CCL2	DDX58	MX1	TNFAIP6	OASL
XAF1	ETV7	TNFAIP6	RTP4	FLJ20035
IFIT2	FCGR1B	RTP4		PML
SOCS1	OAS1	OASL		CXCL10

[a]Corresponds to Figure S3

[b]Corresponds to Factor 16, Figure 1

[c]Corresponds to Factor 20, Figure S5

Figure S1

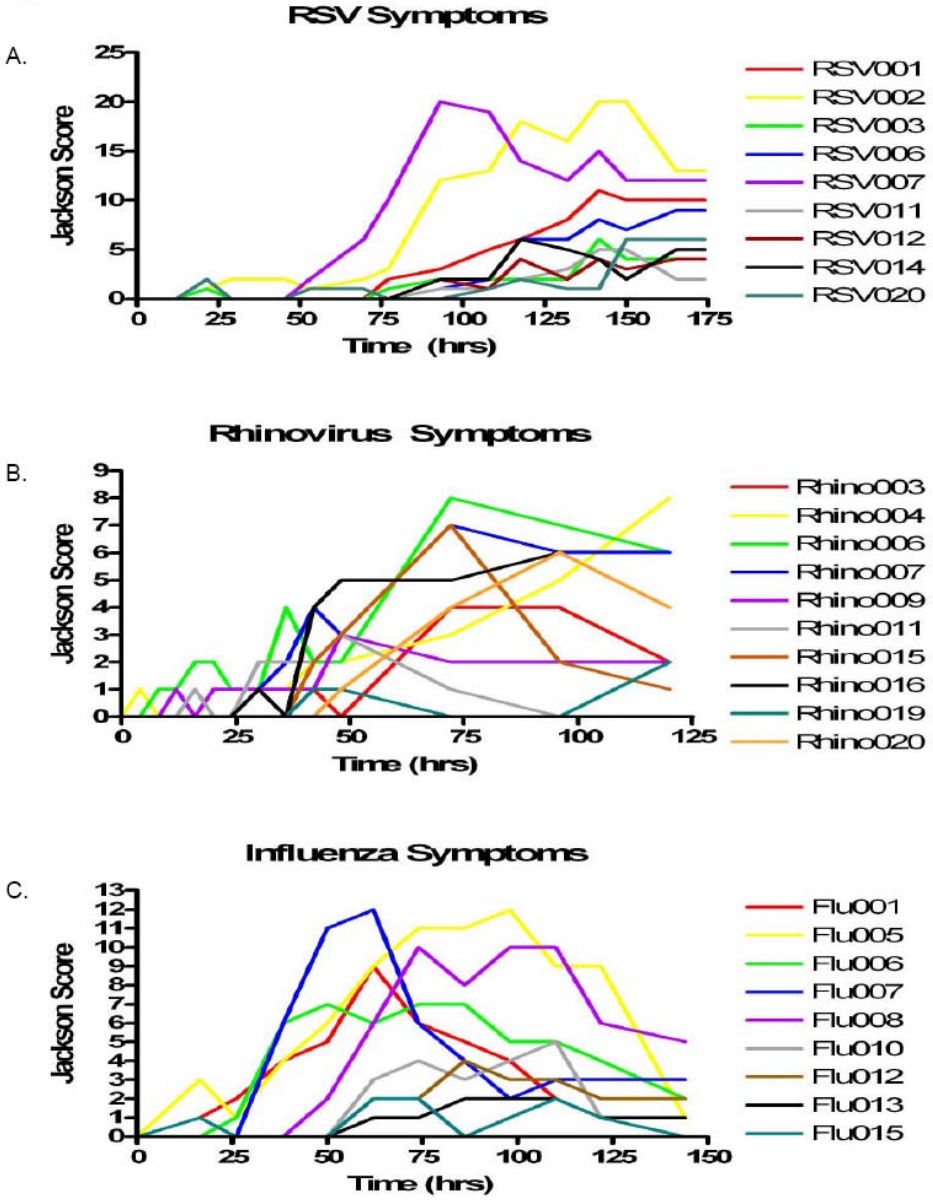

Figure S1.

(A–C) Symptom scores using the Jackson criteria (see Experimental Procedures) across study period (inoculation to end of monitoring) for experimentally infected cohorts. Shown are symptom scores for subjects who developed clinically relevant symptoms only. RSV cohort is shown in (A), HRV cohort is shown in (B), and influenza cohort is shown in (C).

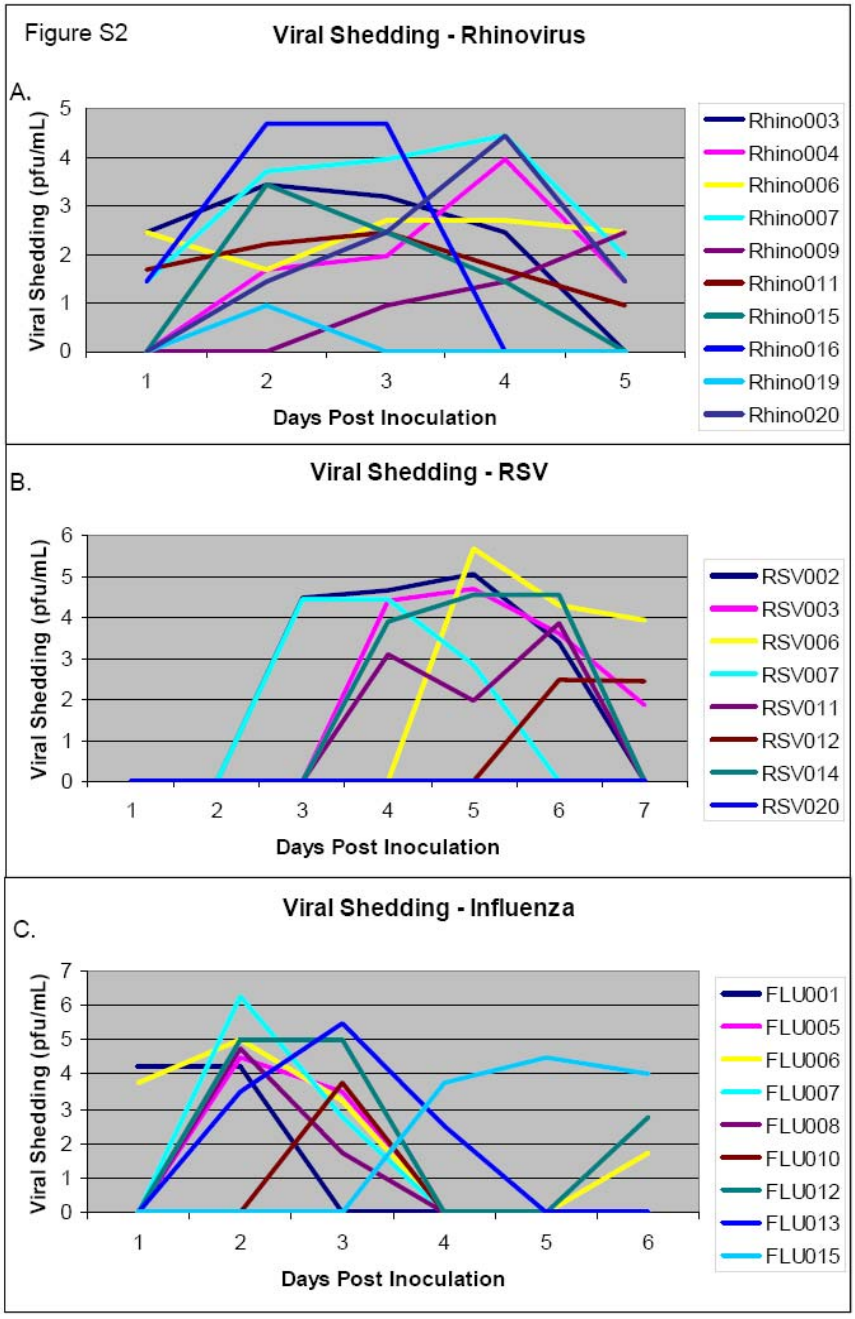

Figure S2.

(A–C) Viral shedding titers as measured by plaque forming units (pfu) per ml. Daily nasal washes were obtained on each subject post inoculation. Viral shedding was determined by viral culture. All symptomatic subjects from the HRV cohort were documented to shed virus (A). All symptomatic subjects except one subject from the RSV cohort were documented to shed virus (B). All symptomatic subjects from the influenza cohort were documented to shed virus (C).

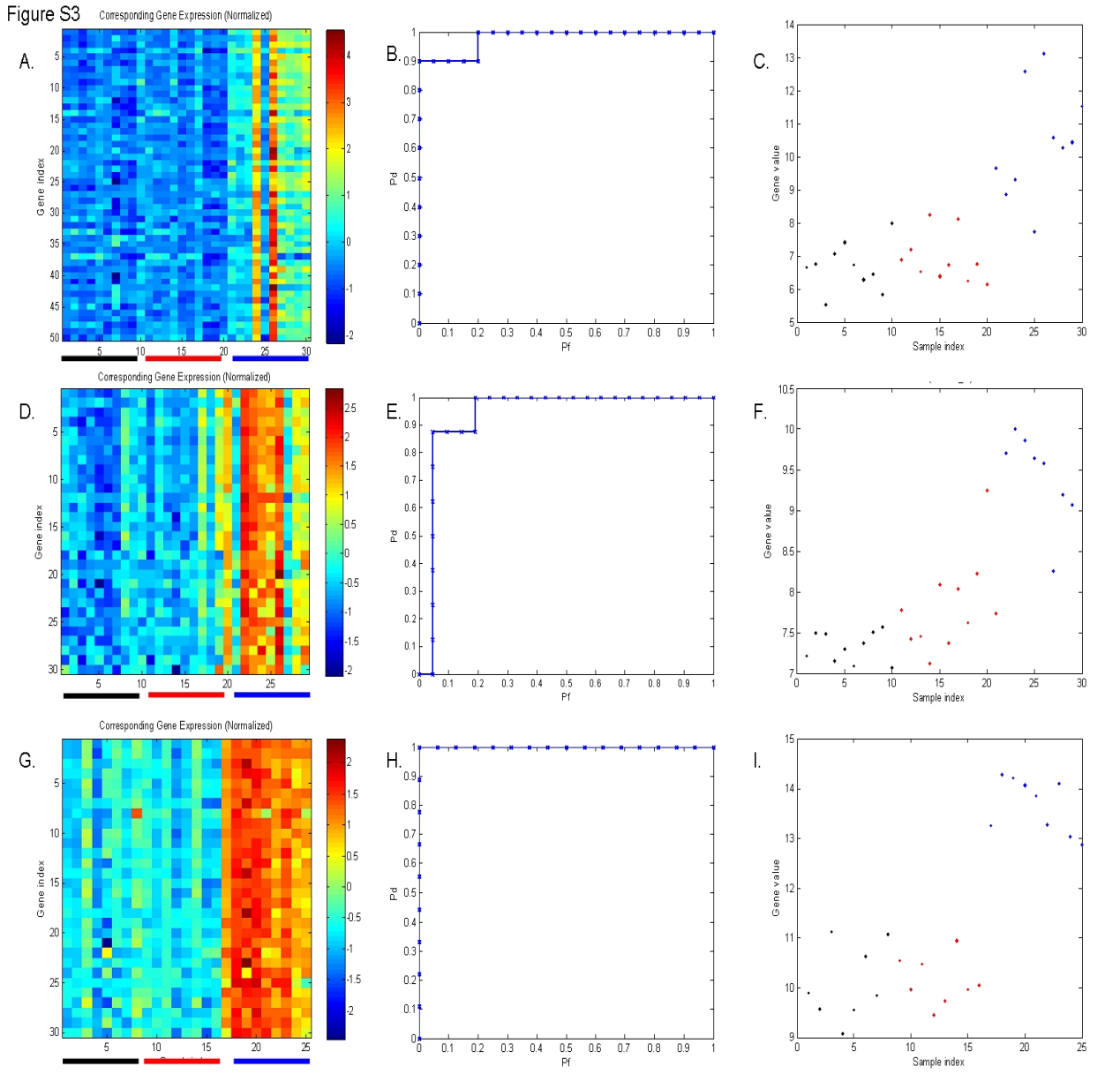

Figure S3. Peripheral Blood Gene Expression Signatures Differentiate Adults with Symptomatic Respiratory Tract Infection from Asymptomatic Adults

(A) Internal cross-validation was performed between each viral challenge dataset. For each individual dataset (HRV, RSV, or influenza A), baseline (preinoculation) gene expression was indistinguishable from the matched timepoint of asymptomatic subjects indicating that the state of "health" defined by gene expression at baseline prior to inoculation was similar to a state of health post inoculation in subjects whom did not develop symptoms. Shown in (A) is a heat map representing relative gene

expression of genes in Factor 6 (HRV). Blue indicates low expression levels and red high expression levels. Samples are arranged along the x axis, corresponding to baseline, asymptomatic and symptomatic and correspond linearly with points in (C). Notably, genes contained in this factor are known to be involved in host defense against viral infection, including interferon signaling (10 genes), host-viral interaction such as MX1, and the 2'-5' oligo (A) synthetase (OAS) gene family (Mashimo et al., 2008; Min and Krug, 2006; Rios et al., 2007) and viral sensing mechanisms (DDX58) (Carvalho et al., 2008). Baseline subjects are indicated with a black bar, asymptomatic with a red bar, and symptomatic with a blue bar, corresponding to the color of subjects in the factor plot shown in (C).
(B) ROC curve for prediction of symptomatic versus asymptomatic subjects in the HRV cohort, generated from probit function utilizing the top 30 genes in Factor 6.
(C) Performance of the gene with the top factor loading score (RSAD2) at discriminating baseline (black), asymptomatic (red), and symptomatic (blue) subjects with experimental HRV infection.
(D) Heat map representing relative gene expression of genes in Factor 20 (RSV). Blue indicates low expression levels and red high expression levels. Samples are arranged along the x axis, corresponding to baseline, asymptomatic and symptomatic and correspond linearly with points in (F). Genes contained in this discriminant factor include interferon-response genes (e.g. IFI44, IFIT1, IFIT3), the OAS family, and viral defense genes such as MX1 and RSAD2 (Proud et al., 2008).
(E) ROC curve for prediction of symptomatic versus asymptomatic subjects in the RSV cohort, generated from probit function utilizing the top 30 genes in Factor 20.
(F) Performance of the gene with the top factor loading score (RTP4) at discriminating baseline (black), asymptomatic (red) and symptomatic (blue) subjects with experimental RSV infection.
(G) Heat map representing relative gene expression of genes in Factor 6 (Influenza). Blue indicates low expression levels and red high expression levels. Samples are arranged along the x axis, corresponding to baseline, asymptomatic, and symptomatic and correspond linearly with points in (I). Genes contained in this discriminant factor include IFN-response genes (e.g. IFI44, IFI44L), SIGLEC1 (a sialoadehesin involved in monocyte response to interferon), the OAS family, and viral defense genes such as MX1 and RSAD2 (Proud et al., 2008).
(H) ROC curve for prediction of symptomatic versus asymptomatic subjects in the influenza cohort, generated from probit function utilizing the top 30 genes in Factor 6.
(I) Performance of the gene with the top factor loading score (ISG15) at discriminating baseline (black), asymptomatic (red) and symptomatic (blue) subjects with experimental RSV infection.

Figure S4

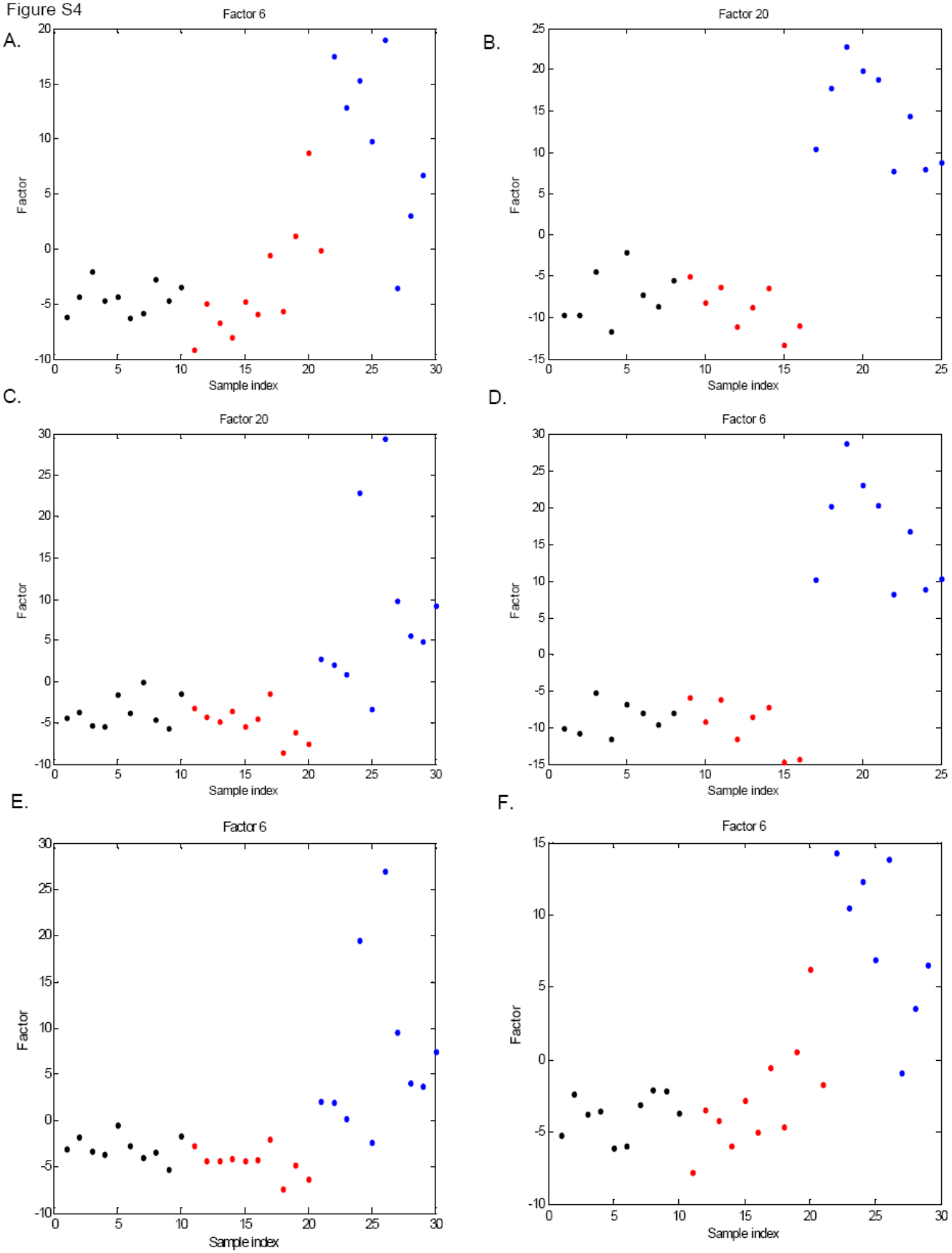
A.
Factor 6
Factor
Sample index
B.
Factor 20
Factor
Sample index
C.
Factor 20
Factor
Sample index
D.
Factor 6
Factor
Sample index
E.
Factor 6
Factor
Sample index
F.
Factor 6
Factor
Sample index

Figure S4. Cross-Validation for Predicting Symptom Class (Asymptomatic versus Symptomatic) for Individual Respiratory Viral Illnesses Based on Training on Separate Viral Challenge Data

(A) Prediction of symptomatic (blue) subjects with RSV infection versus asymptomatic (red) subjects utilizing Factor 6 derived from HRV. Prediction accuracy across panels is high, indicating dominance of a pan-respiratory viral response.

(B) Prediction of symptomatic (blue) subjects with influenza infection versus asymptomatic (red) subjects using Factor 6 derived from HRV.

(C) Prediction of symptomatic (blue) subjects with HRV infection versus asymptomatic (red) subjects utilizing Factor 20 derived from RSV.

(D) Prediction of symptomatic (blue) subjects with influenza infection versus asymptomatic (red) subjects utilizing Factor 20 derived from RSV.

(E) Prediction of symptomatic (blue) subjects with HRV infection versus asymptomatic (red) subjects utilizing Factor 6 derived from influenza.

(F) Prediction of symptomatic (blue) subjects with RSV infection versus asymptomatic (red) subjects utilizing Factor 6 derived from influenza virus. Factor numbers are arbitrary; thus, Factor 6 from HRV data and Factor 6 from influenza virus data are not the same factor.

Figure S5

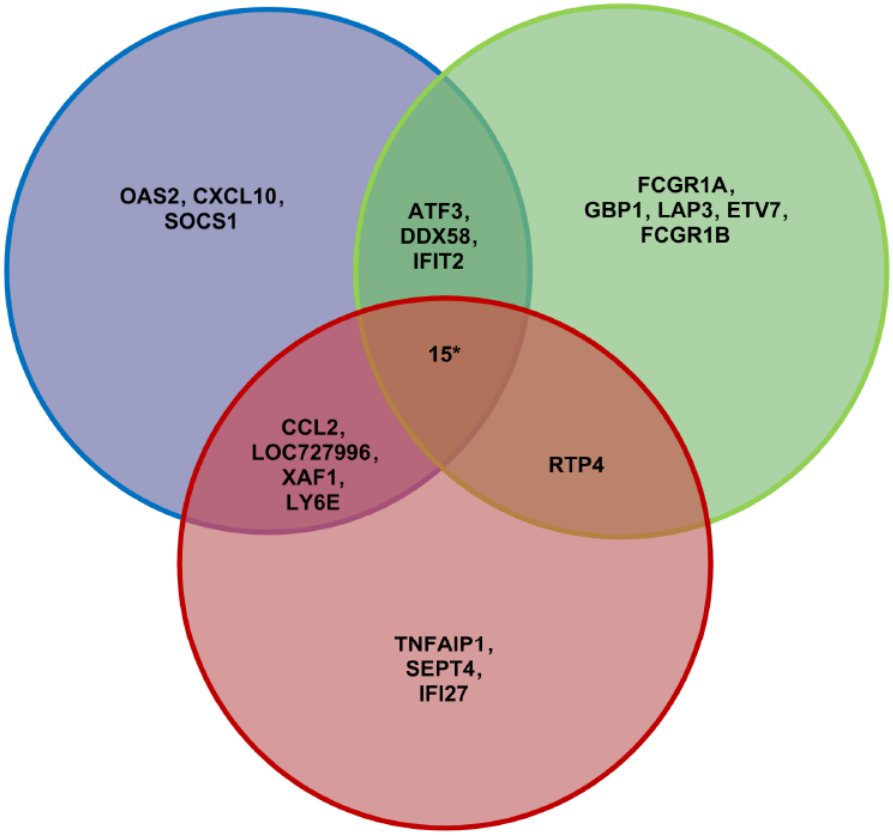

Figure S5. Venn Diagram Illustration Indicating Overlap between Genes in the Three Individual Viral Factors

Twenty-seven unique genes characterized the acute respiratory viral illness factor. Fifteen genes were common to the three individual virus factors and are listed below the diagram. Three genes were present in the HRV factor alone (OAS2, CXCL10, and SOCS1). Five genes were present in the RSV factor alone (FCGR1A, GBP1, LAP3, ETV7, and FCGR1B). Three genes were present in the influenza factor alone (TNFAIP1, SEPT4, and IFI27). Notably, the only genes present in the individual viral factors that were not represented in the pan-respiratory viral factor were SOCS1 (HRV) and FCGR1A, GBP1, LAP3, ETV7, and FCGR1B (RSV) and TNFAIP1 (influenza). A complete list of genes represented in each factor is shown in Table S1. Thus, a generalized viral response signature dominates over individual viral responses at time of peak symptoms in experimentally induced respiratory viral infection.

Figure S6

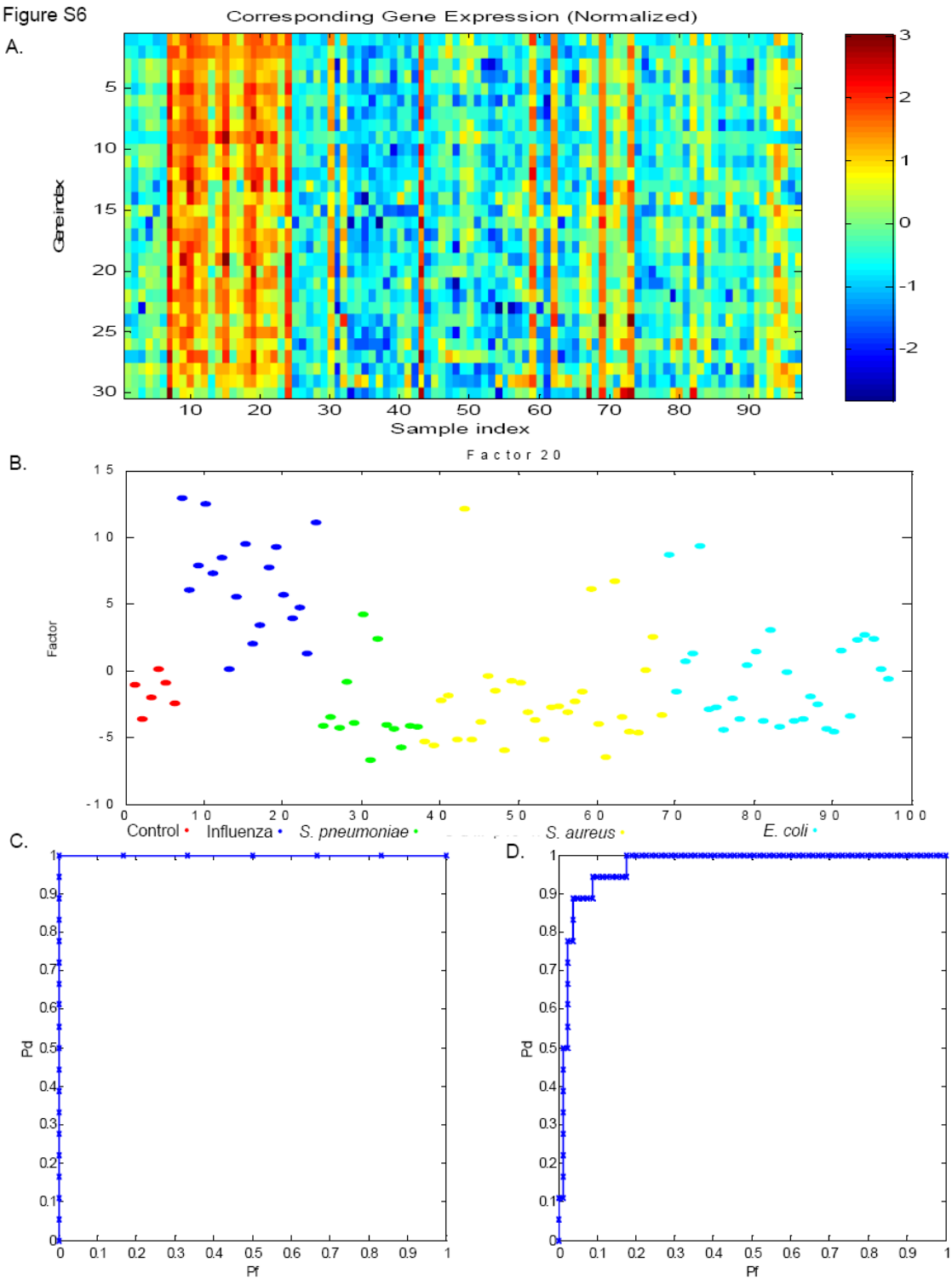
A.
Corresponding Gene Expression (Normalized)
Gene index
Sample index
B.
Factor 20
Factor
Control
Influenza
S. pneumoniae
S. aureus
E. coli
C.
Pd
Pf
D.
Pd
Pf

Figure S6.

(A) Sparse latent factor regression analysis applied to an existing dataset of PBMC expression in microbiologically confirmed infections (influenza A, *S. pneumoniae*, *S. aureus, E. coli*) can discriminate between individuals with influenza A infection and healthy controls, as well as individuals with bacterial infection (red = uninfected, blue = influenza A, green = *S. pneumoniae,* light blue = *E. coli*, yellow = *S. aureus*).

(B) Leave one out cross-validation correctly identifies 100% of individuals with influenza A infection versus no infection (0/24 misclassified).

(C) Leave one out cross-validation correctly identifies 93% (7/97) individuals with influenza A infection versus bacterial infection. Pd = probability of detection, Pf = probability of false discovery.

Figure S7

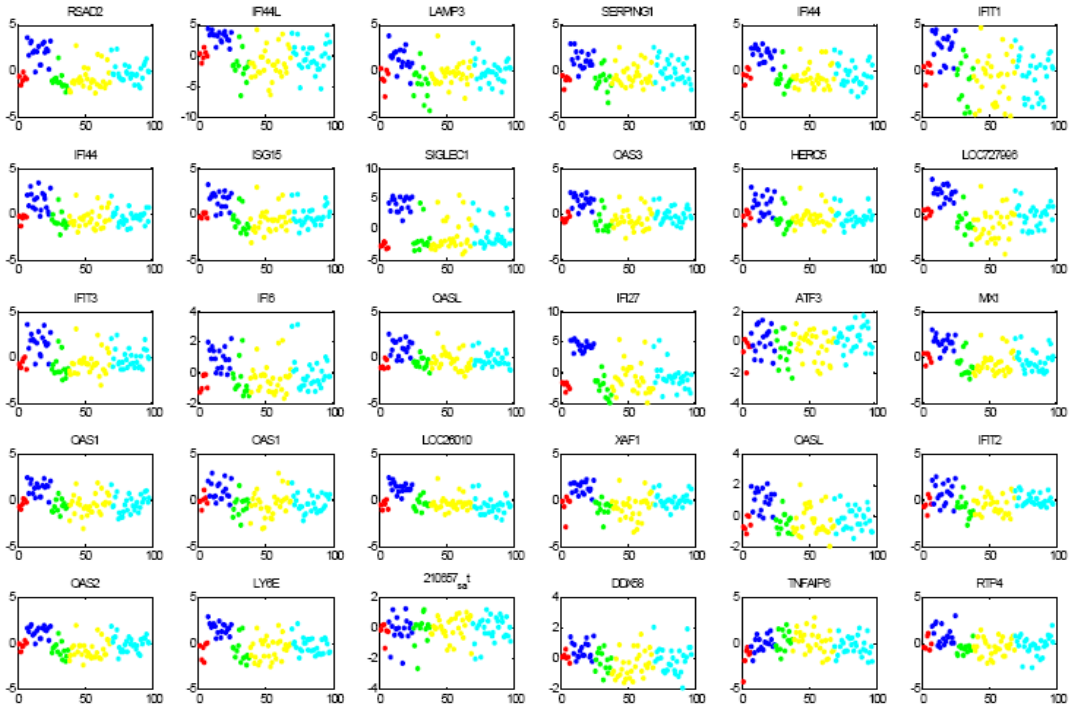

Figure S7. Predictive Performance of Each Gene Contained in the Probit Function Generated from the Acute Respiratory Viral Factor to Predict Pathogen Class in the Independent Data Set

Gene name is shown at the top of each plot. Subjects are color-coded by infecting pathogen (red = uninfected, blue = influenza A, green = *S. pneumoniae*, light blue = *E. coli*, yellow = *S. aureus*).

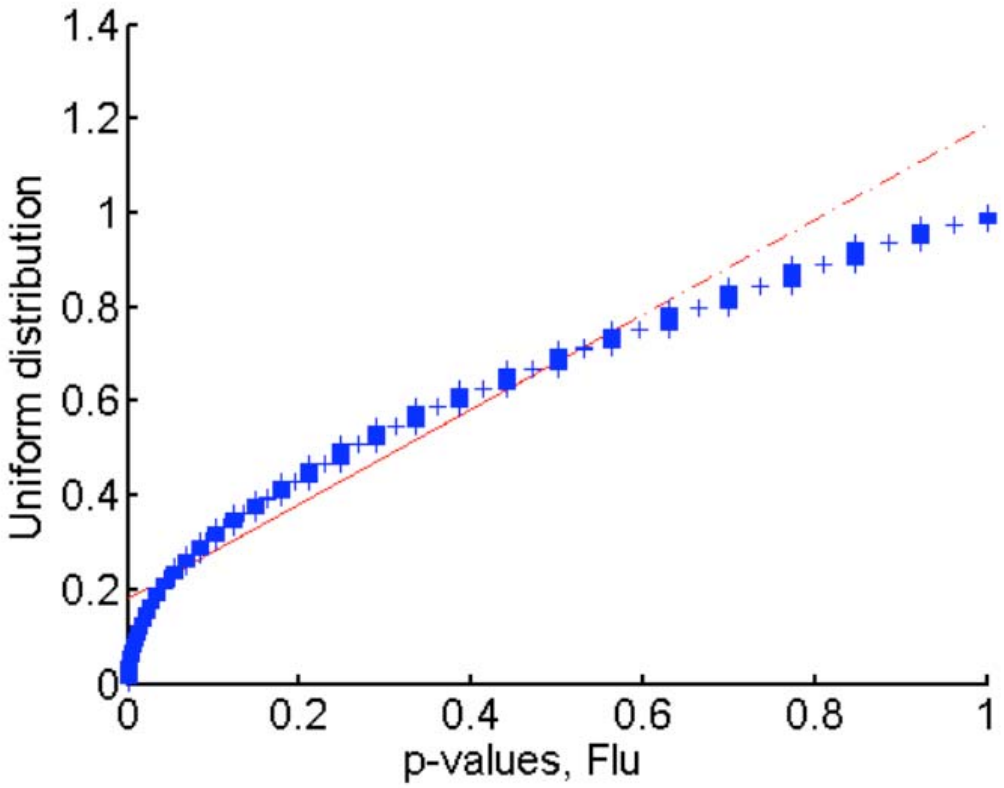

Figure S8. A Quartile Plot Comparing a Uniform Distribution to the P-Values Obtained from Testing Each Gene

P-values were generated with a nonparametric Kruskal-Wallis test for significant difference in location between symptomatic and asymptomatic subjects at time T in the influenza cohort. The curl near zero indicates a predominance of low p-values in the data, although there are no individual genes that retain significance after correction for multiple hypotheses.

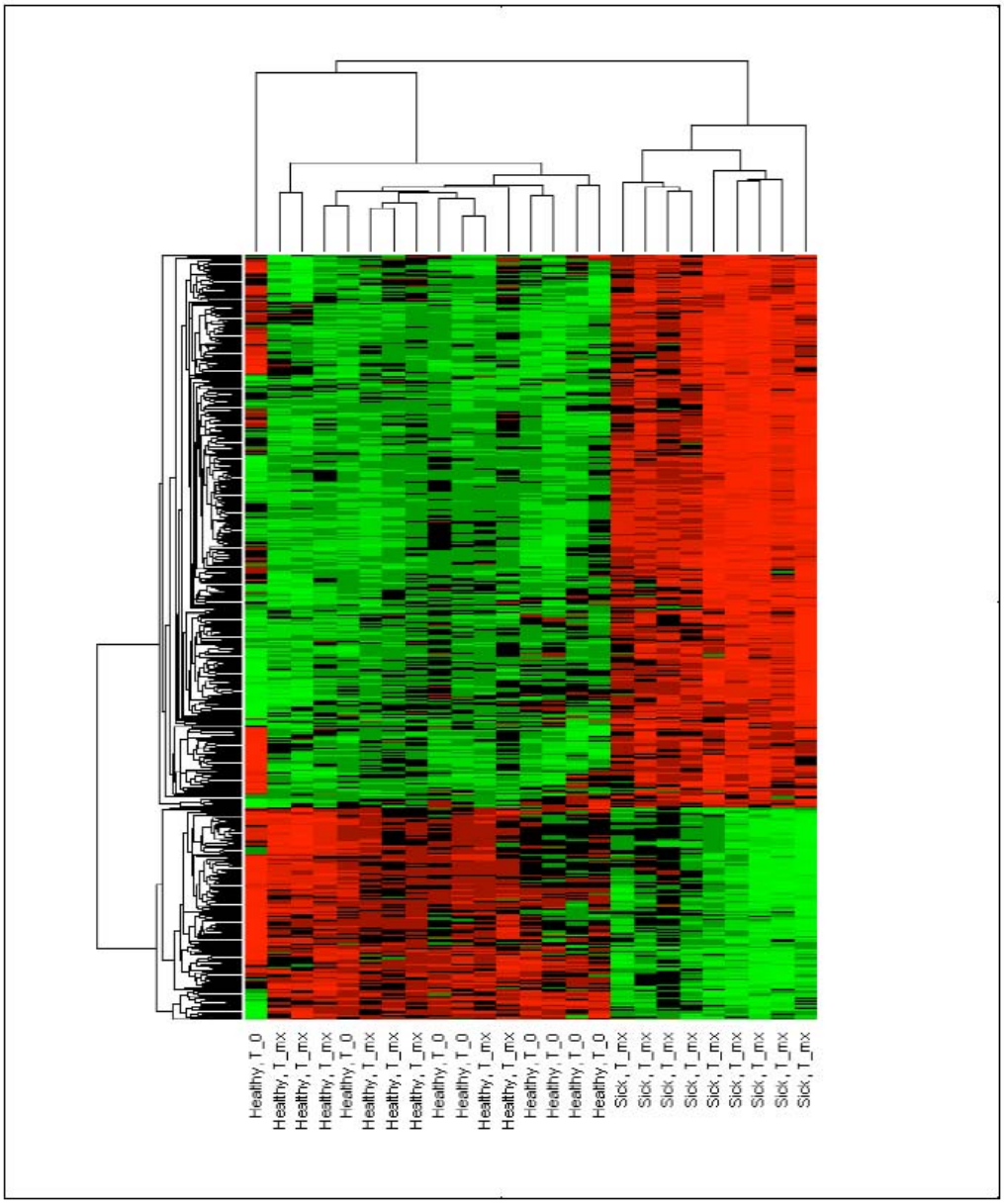

Figure S9. Clustering, Flu

Clustering of the flu cohort on only those genes that are identified as being differentially expressed ($p < 0.001$) between sick and healthy subjects at t_{max}. The subjects split according to health, with no apparent distinction between healthy patients at different blood draw times.

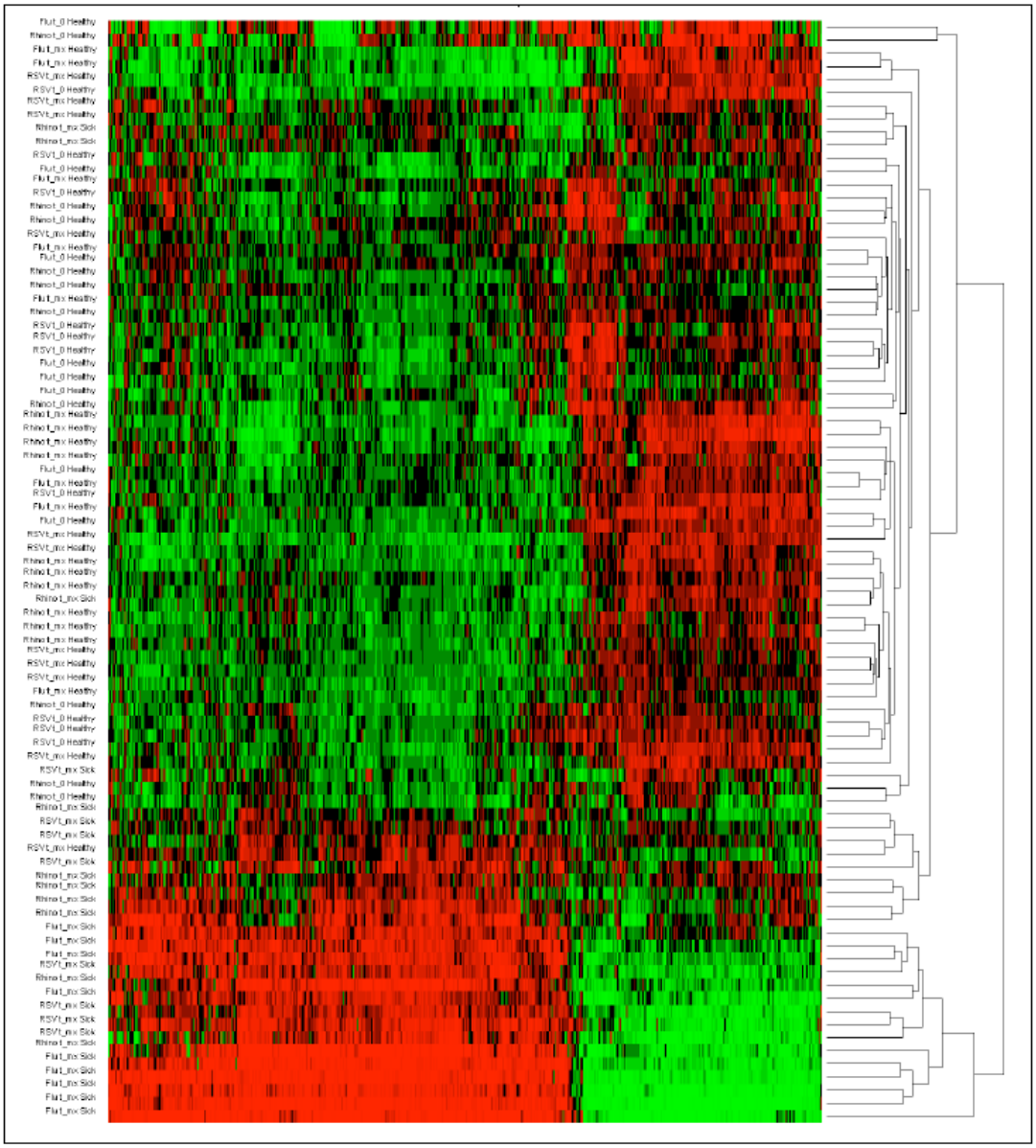

Figure S10. Clustering, All Samples

Clustering on only those genes significantly associated ($p < 0.05/22{,}215$) with t_{max}, sick versus healthy or interactions between health status and one of the specific disease types (ANOVA). Healthy and sick patients are generally grouped together, with no discernable difference between healthy patients measured at t_0 and healthy patients measured at t_{max}. The algorithm has grouped the sick patients into two separate clusters, one of which contains all of the flu patients, indicating the possibility that flu may be distinguishable from other viral infections based on host response at the time of maximum symptoms.