Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

04/15/2020

_____ _____

Chunhui Gu Date

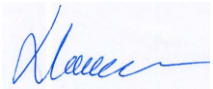
Whole-blood transcriptional signatures of different severity profiles in patients with influenza A H1N1 infection

By

Chunhui Gu

Master of Science in Public Health

Department of Biostatistics and Bioinformatics

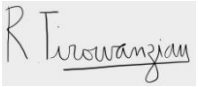


Hao Wu, Ph.D.

Committee Chair

Diego Mauricio Moncada Giraldo, Ph.D.

Committee Member



Rabindra Tirouvanziam, Ph.D.

Committee Member

Whole-blood transcriptional signatures of different severity in influenza A H1N1 infection

By

Chunhui Gu

MBBS, Fudan University, 2018

Thesis Committee Chair: Hao Wu, Ph.D.

An abstract of

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Department of Biostatistics and Bioinformatics

2020

Abstract

Whole-blood transcriptional signatures of different severity profiles in patients with influenza A H1N1 infection

By Chunhui Gu

Announcement: This is a re-analysis research article for a public dataset (GSE111368).

Background: As influenza remains one of the major threats worldwide because of its high mutational and recombination rates, vaccines and antivirals continue to be developed that help the body fight against infection through adjusting host gene expression for as many strains of the virus as possible.

Objective: The primary purpose of this study is to investigate which genes or gene sets are related to the development of a high severity profile in patients with influenza infection.

Methods: Normalized microarray data were obtained from a prior publicly available influenza study (Dunning et al. Nat Immunol, 2018). Differential expression analysis was accomplished by R version 3.6.1 and "limma" package version 3.40.6. Gene set enrichment analysis (GSEA) was conducted by using Molecular Signatures Database (MsigDB). CibersortX, an enhanced digital dissection technique was used to generate cell-specific gene expression patterns.

Results: The differential expression analysis (DEA) showed that 48 (236), 158 (323), and 293 (425) genes were differentially down-regulated (up-regulated) in severity I, severity II, and severity III categories of influenza patients, respectively, as compared to healthy controls (the changes of expression in severity categories compared with healthy controls were called contrasts for short). In total, there were 824 distinct differentially expressed genes (DEGs). 138 of the 824 DEGs were considered to have obvious different fold-changes under different contrasts and the difference in fold-change mainly came from the comparison between high severity contrast and low severity contrast. Several gene sets were identified based on gene set enrichment analysis and over-representation analysis.

Discussion: The analysis of overlaps between enriched gene sets of different severity profiles of patients with influenza (compared to healthy controls) demonstrated that autoimmune mechanisms could play an important role in the development of a high severity profile. There remained around half of 138 DEGs that could not be attributed to any gene set in target collections. A custom set can be built from those genes as possible gene set describing the development of influenza A pathogenesis. The cell-specific transcriptional signature generated by CibersortX suggested that neutrophils and monocytes represent the main cell-types in which the effect of those genes and gene sets takes place.

Whole-blood transcriptional signatures of different severity profiles in patients with influenza A H1N1 infection

By

Chunhui Gu

MBBS, Fudan University, 2018

Thesis Committee Chair: Hao Wu, Ph.D.

An abstract of

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Department of Biostatistics and Bioinformatics

2020

Acknowledgments

I would like to thank Dr. Rabindra Tirouvanziam for providing such a great opportunity for me to conduct this thesis and full discretion during the writing of the thesis. It was my honor to do my thesis with Dr. Tirouvanziam's lab where several previous students from the Department of Biostatistics and Bioinformatics had shown significant growth within their thesis projects and manuscripts.

I would like to thank Dr. Diego Mauricio Moncada Giraldo, who acted as my mentor, for teaching me all the bioinformatics and biology knowledge needed to write this thesis, and for supporting me when I felt frustrated. I appreciated his endless patience in explaining even very basic biological knowledge to me.

I would like to thank Dr. Hao Wu for not only providing valuable advice in writing the thesis but also guiding me on how to work as a professional in this area: to be responsible, discreet, and enterprising.

Finally, I would like to thank my family and girlfriend, for supporting me throughout the two years of study abroad. Writing a thesis is never an easy task, but I have learned a lot during this process and doing so, prepared myself better for the next stage of my career.

Table of Contents

Introduction

Influenza disease caused by influenza virus infection is a global disease and was estimated to cause 290,000 to 650,000 influenza-related respiratory deaths every year during 2019-2030 [1]. Influenza viruses present as four types: A, B, C, and D [2]. Influenza A virus (**IAV**) is known for causing epidemic or even pandemic influenza outbreaks in humans. The most recent pandemic influenza (influenza A pdm09) occurred in 2009, which led to at least 6,670 deaths and more than 526,000 infections globally [3].

The high mutational and recombination rates of influenza viruses especially IAV, are a major reason for the lack of long-lasting individual and herd immunity, causing seasonal epidemics or worldwide pandemics. IAV can be classified into different subtypes based on two surface glycoproteins: hemagglutinin (**HA**) and neuraminidase (**NA**) [4], which control virus entry into respiratory cell [5] and release of virus offspring from infected cells. A typical HA protein consists of three parts – a head with Receptor Binding Site, antigenic determinants, and a stalk region [6]. The high variability of the head region of HA complicates the design vaccines aiming to induce the production of antibodies targeting HA, because the correct prediction of which among circulating strains are going to become prevalent in the future is very difficult [7]. Although universal IAV vaccines are theoretically possible because of the relatively high conservation of RBS and stalk regions between strains, cross-protection provided by those vaccines is too weak to in practice [8] [9]. The stability of NA makes it another target for universal IAV drugs. Neuraminidase inhibitors (NAIs) can block the enzymatic site of the NA protein involves in the release of virus offspring from infected cells. Oseltamivir and zanamivir are two

NAI drugs that are widely used currently. Unfortunately, the occurrence of variants resistant to oseltamivir largely undermines its effect [10]. Despite the fact that only a few IAV strands have shown resistance to zanamivir, the debate about its ability in treating IAV remains unsettled [2] [11] [12].

While much effort has focused on how to prevent and treat IAV infection by targeting the virus itself, the effect of host factors in determining the course of influenza infection was less known, at least until recently. Arguably, the high variability in observed clinical outcomes must have something to do with host factors, considering that most influenza infections are followed by mild symptoms, while a few cases lead to severe outcomes [4]. The impact on hospitalization of co-morbidities (such as obesity) and sociodemographic factors (such as a low education level) have been studied during the 2009 influenza (H1N1) pandemics [13]. However, more than 1/3 of all hospitalizations could not be explained by the identified host factors, leaving the possibility that other, as yet unknown, host factors may contribute to the progression of influenza infection [14]. Genetic risk factors such as single nucleotide polymorphisms (**SNPs**) may be at play, some with potential impact on the severity of influenza symptoms such as the rs12252 SNP in the IFITM3 gene [15]. However, no significant host SNP was identified that significantly correlated to poor prognosis in the 2009 influenza A (H1N1) pandemics [14].

Explaining the largely fruitless efforts so far to pin differences in influenza infection on one or two genes, it is highly likely that many host genes are involved as a network in the response to influenza infection. Several studies have revealed the transcriptional signature of IVA infection [16] [17] and use it for diagnosis [18]. But those studies only involved a small sample size and didn't explore gene expression patterns in patients with

different severity profiles in response to IAV infection. Recently, a study with a large enough sample size demonstrated that the transcriptional signatures in patients with different severity profiles are largely different. This leads to significantly different activation of some modular biological functions compared with healthy controls, including a down-regulation in interferon-related transcripts and an up-regulation in the inflammation module in the group with the highest severity [19]. However, this study did not consider that the different signatures associated with the various severity profiles could relate to confounders such as bacterial co-infection which accounts for about 11% to 35% of influenza infection [20]. Finding differentially expressed genes linked to influenza infection is not sufficient as one still needs to know *why* those genes are differentially expressed in order to accurately identify those that can help the host fight the virus. For example, two different explanations can be advanced for down-regulated genes. One is that IAV can employ mechanisms to reduce expressions of protective host genes and production of host proteins ("host shutoff"), which contributes to immune evasion of the virus, redirecting resources to the production of viral proteins [21]. Alternatively certain genes may be down-regulated to reallocate resources toward the production of antiviral proteins. Using new gene editing technologies to specifically up-regulate those down-regulated genes could help the immune system in the battle with viruses in the first case while having the opposite effect if the truth is the alternative explanation.

The primary purpose of this study was to investigate which genes may be responsible for high disease severity in patients with influenza after correcting for confounding factors.

Identifying those genes may lead to prophylactic measures to help prepare the immune system for IAV infection and avoid developing severe symptoms.

Methods

Study dataset

Data used in this study originated from a previously published study [19] funded by the Mechanism of Severe Acute Influenza Consortium (MOSAIC), for which 109 influenza patients and 130 healthy controls were recruited during the 2010-2011 time span. Samples from three timepoints were obtained: T1 (at recruitment), T2 (48 hours after T1), and T3 (≥ 4 weeks after T1). The severity of influenza symptoms was categorized at T1 and T2 according to three levels: level 1, no considerable respiratory compromise, and blood oxygen saturation ≥ 93% without using additional oxygen supply other than room air; level 2, oxygen saturation is ≤ 93% with or without additional oxygen supply; level 3, respiratory compromise with invasive mechanical ventilation. More details about grading criteria can be found in the original paper.

Microarray data processing and normalization

Since raw data (idat file) were not provided by the authors despite repeated demands, the normalized data were used as a result. As quoted in the original paper, the following processes were applied to obtain normalized data from raw data: "*Raw microarray data were processed using GeneSpring GX version 12.5 (Agilent Technologies). Following background subtraction, each probe was attributed a flag to denote its signal-intensity-detection P-value. Filtering on flags removed probe sets that did not result in a 'present' call in at least 10% of the samples, where the 'present' lower cut-off was 0.99. Signal*

values were then set to a threshold level of 10, were log2-transformed and were per chip normalized using a 75th percentile-shift algorithm. Each gene was normalized by dividing each mRNA transcript by the median intensity of all samples" [19]. Details about how normalized data was generalized are described in this original paper. The mapping from the Illumina manufacturer identifiers (ILMN) to gene symbols was completed by using a revised version of annotation illuminaHumanv4SYMBOLREANNOTATED [22] in illuminaHumanv4.db database. Average values were calculated for each distinct gene that has duplicate probes matched to it.

Selection criteria of samples

The primary purpose of this study was to investigate genes or gene sets related to different severity profiles in patients with influenza A H1N1 (we will use influenza A for short in the rest of this report) after adjustment for any confounders. Since there was no severity information provided in T3, data at T3 were not included in the following analysis. One important goal of this study is to distinguish genes differentially expressed because of influenza A infection rather than bacterial infection. Bacterial infection status, however, was not available for all influenza A patients, and we used samples from healthy controls and patients with valid bacterial infection status. One sample was further excluded as an outlier after analysis by PCA (**Supplementary Fig. 1**). In all, 227 samples from 61 patients and 130 healthy controls were used as the final dataset.

Differential expression analysis design

Differential expression analysis (**DEA**) was accomplished by R version 3.6.1 and "limma" version 3.40.6 (details about how to construct the data for DEA can be found in

the **Supplementary methods**). Genes that are related to a higher severity profile were obtained by analyzing the differential expression genes (**DEGs**) from the following contrasts in the model: severity I group vs. healthy control group, severity II group vs. the healthy control group and severity III groups vs. the healthy control group (**Equation 1**). Those three contrasts were designated S1-HC contrast, S2-HC contrast, and S3-HC contrast in the rest of this report for simplicity. False discovery rate adjusted p-value (**FDR adjusted p-value or q-value**) was used to control for multiple testing per the Benjamini-Hochberg method. A gene was considered as differentially expressed if its adjusted p-value was less than 0.1 and with an absolute log two-fold-change greater than 1. Heatmaps of DEGs were created using hierarchical clustering with the "complete" method by "pheatmap" package version 1.0.12 [23].

Gene set enrichment analysis

Gene set enrichment analysis (**GSEA**) was a useful technique for finding meaningful DEGs by classifying genes into annotated groups related to specific biology processes [24]. We conducted GSEA by using the "fgsea" package, which implements cumulative GSEA-statistic calculation [25] along with the Molecular Signatures Database (MsigDB) [26]. An ordered list (ordered by log2-fold-change) of all the gene names and corresponding log2-fold-changes from DEA was used for each contrast in GSEA. To identify gene signatures in the dataset, overlaps with four collections of gene sets were used: HALLMARK, Biocarta, Kyoto Encyclopedia of Genes and Genomes (KEGG), and REACTOME. The minimum size of a gene set to test was set at 15 for reliable results. The maximum size of a gene set to test was set as large enough (5,000) to include all gene sets beyond the minimum requirement. The number of permutations to get adjusted

p-values was 50,000 for every candidate gene set for all three contrasts. The normalized enrichment score (**NES**) was obtained by divided the enrichment score (**ES**) by the mean enrichment of random samples of the same size. For each contrast, we took the top 20 sets ordered by absolute NES after filtered by FDR adjusted p-value of 0.25 and analyzed their overlap condition. Top 40 and top 60 options were also used with the same procedure as we just described for the top 20 option.

Imputing cell fractions and high-resolution cell-specific expression by CibersortX

CibersortX [27], a web-based enhanced digital cytometry technique was used to infer cell fractions and cell-specific gene expression patterns for each sample. The cell fractions were imputed with the following parameters: no batch correction and no quantile normalization since no batch information was available and data were already quantile normalized. The high-resolution cell expression mode at this time only accepts less than 1,000 genes due to intense computational demand. A list of DEGs was used as a subset file. The data were merged into 10 major cell subsets by the signature matrix LM22. High-resolution cell expression was imputed with the following parameters: no batch correction, no quantile normalization, and with default window size for deconvolution.

Results

Most genomics studies about influenza A focus on the difference in whole transcriptional signature between subjects versus healthy people without considering their different severities. In this study, we will concentrate on how genes are differentially expressed in different severity conditions after adjusting for possible cofounders. The data used in this study was a public dataset with a large enough sample size funded by MOSAIC.

"Limma" package in R was used to build a linear model to explore genes related to a high severity after adjusting for other necessary covariates.

Demographic metadata covariance in different severity group

To identify differences in gene signatures between different severity stages, it is essential to adjust for possible confounders and correlation structure. The demographic table was constructed to check the imbalance distribution between different severity groups for demographic variables using the chi-square test or ANOVA test. The dataset used in this study contained 227 samples from 61 influenza patients and 130 healthy controls (**Supplementary Table 1**). Each of the 130 healthy controls only contributed one sample; 25 influenza patients contributed only one sample and 36 patients contributed 2 samples at two different time points, which introduced a correlation between samples (this will be discussed more later). The 227 samples were classified into 4 categories: 130 samples of healthy control, 37 samples of severity I, 25 samples of severity II, and 35 samples of severity III. Age ($p < .001$), day of illness at sampling ($p < .001$), pregnancy ($p = 0.003$), comorbidities ($p <.001$), and bacterial infection status ($p < .001$) were considered as significantly different among those four groups and were treated as covariates to be adjusted. Meanwhile, since ethnicity ($p = 0.651$) and sex ($p = 0.986$) in different groups were quite balanced, they were not adjusted in the model (**Table 1**).

Principal component analysis for different severity group

To obtain gene signatures from the microarray data, we used the improved probe-to-gene annotation to translate the probe manufacture ID to the gene symbol, based on which 18,651 genes were mapped from 18,974 probes. Average values were calculated for each

distinct gene, and 12,828 distinct genes were obtained. Principal component analysis (**PCA**) of those 12,828 genes showed that the whole blood signature distinguished samples from healthy controls and influenza patients by the first two components (21.3% and 6.1%) (**Fig. 1A**). Among different severity groups, the boundary between severity I and severity II groups was blurred, but there was a relatively clear boundary between severity III group and the other two severity groups, which suggested that a representative group of genes exists that can be used to differentiate a high severity stage from a low severity stage in influenza A infection.

Differential expression analysis

Since some of the samples in the influenza group came from the same patients, the correlation between technical replicates was estimated by the "duplicateCorrelation" function [28] in "limma" package, which returned an average correlation on the atanh-transformed scale of 0.259 used to adjust the batch effect from the repeated measurement from the same patients. The DEA showed that there were 48 (236), 158 (323), and 293 (425) genes were differentially down-regulated (up-regulated) respectively in the S1-HC contrast, S2-HC contrast, and S3-HC contrast, which in total were 344 (495) distinct genes (**Fig. 1B**). Moreover, not only there were more up-regulated genes, but their fold-change levels are also considerably higher than down-regulated genes; due to the large sample size, almost all genes had an adjusted p-value less than 0.1 (**Fig. 1C**).

There is a total of 824 distinct DEGs in any of those contrasts. The normalized intensity of those distinct 824 DEGs in three influenza severity groups and healthy control group revealed different transcription signatures. Differences between severity III group and the other two severity groups were also apparent for some clusters of genes (**Supplementary**

Fig. 2). 138 of the 824 DEGs were considered to have different fold-changes under different contrasts (a DEG was included if its fold-changes of any of three contrasts is one time larger than another contrast). We observed that the difference in fold-change mainly came from the comparison between S3-HC contrast and low severity contrast (S1-HC or/and S2-HC) (**Fig. 1D, Supplementary Fig. 3**).

Gene set enrichment analysis

GESA showed a similar result to DEA. The Venn diagram displayed very similar patterns when we used the top 20, 40, and 60 gene sets for each comparison in the three contrasts. The number of gene sets in S1-HC and S2-HC just changed a little bit after switching from top 20 to top 40 because there were less than 40 gene sets filtered by FDR p-value of 0.25, and almost the same from top 40 to top 60 (**Fig. 6**). As a result, we kept using the top 40 because it explored all the important gene sets in two higher severity contrast but did not introduce unimportant gene sets from the severity I – HC contrast. 13 gene sets were enriched in all those three contrasts. 4 of those 13 gene sets were upregulated on all three contrasts and were all related to interferon signaling. 9 of 13 gene sets were down-regulated in S1-HC contrast but up-regulated in higher severity contrasts (S2-HC and S3-HC). 5 gene sets were exclusively enriched in the S3-HC contrast, of which 4 came from REACTOME (axon guidance, developmental biology, metabolism of amino acids and derivatives, and translation) and 1 came from HALLMARK (allograft rejection).

Imputing cell fractions and high-resolution cell-specific expression by CibersortX

As the severity of disease increases, the proportion of T cells decreased while the proportion of B cells increased (**Supplementary Fig. 4**). t-SNE plots showed the cluster

structure of samples with different severity status for different cell types. In none of those plots, severity I and severity II had unclear boundaries and were even mixed with healthy controls in some cell types (B cells and dendritic cells). For severity III patients, however, visible boundaries were seen with healthy controls except for dendritic cells (**Fig. 3B, Supplementary Fig. 6**). Severity III showed a blurred boundary between severity I and severity II group for neutrophils and monocytes and they also had the most DEGs showing expression differences between healthy controls and disease samples (**Fig. 3A, Supplementary Fig. 5**).

Discussion

This study identifies genes or gene sets related to the severity of disease in patients with influenza A infection. We distinguished genes and gene sets as candidates that may serve as candidates for interventions with small molecules or gene therapy.

We filtered 138 DEGs that were differentially regulated under different severity stages, just a few of them were in any of those 52 enriched gene sets in GSEA that had an FDR adjusted p-value less than 0.25. An over-representation analysis (**ORA**) demonstrated that about half of those genes fall into gene sets similar in GESA **(Supplementary Table 1)** and still more than half of those genes that were not significant in any of gene sets in those 4 gene set collections. Although gene-level DEA is not as stable as gene-set-level GSEA, genes with differences in at least one log2-fold-changes showed signatures of neutrophil activation and degranulation as contrasting differences between different severity groups (**Fig. 1D**). Accordingly, we suggest using this gene set to set up a new signature for severe cases in influenza infections using the 138 DEGs that differentially

expressed under high and low severity stages. In the future, we can use this custom gene set to conduct GESAs for data from other studies to further explore that whether those genes could be used as a pathway to demonstrate how our body fight with influenza A infection on the high severity stage.

While we have targeted some genes that may be associated with influenza A infection severity, we cannot prove that whether differential expression of those genes is the reason why patients fall into a high severity stage or the result from the body's response to influenza A infection. To address this specific issue, future studies are needed that will obtain blood samples at different time points from a sufficient number of patients after symptom onset. By observing the change in severity over time, we can group them into two major outcome groups: relief group and aggravation group. By analyzing the trend of gene expression in different groups and find genes that have different expression regulation profiles, we could be more confident in concluding those genes related to the development of infection after adjusting for possible confounders.

Although the dataset used in this study was not designed for this purpose, we can still make reasonable inferences from GSEA results and rule out genes and enriched gene sets that are not likely to be the reason why a high severity stage is reached. The leading edges of the four gene sets from REACTOME collection enriched exclusively in the S3-HC contrast showed that the majority of genes shared between the four gene sets were Myc targets (**Supplementary Fig. 7**). Those genes are involved in cell proliferation, apoptosis, and metabolism and seem to relate to the recovery process of the body [29]. Besides, many up-regulated gene sets related to autoimmune progress such as allograft,

graft versus host disease (GVHD), and asthma suggests that the autoimmune mechanism could play an important role in the development of symptoms [30].

The CibersortX tool allowed us to investigate cell-type-specific gene signatures. However, at this time, it only generated matrices for a gene list of a certain length for computation consideration. If the transcriptional signature of all the genes in the original normalized matrix is available for each specific cell type, we can use that just as a normalized matrix comes from a microarray or RNA-Seq to conduct DEA. Besides, the imputed cell-specific expression matrices were limited in the resolution. Some mathematical transformations needed to be used to magnify the contrast before the plotting of the heatmap of cell-type-specific transcriptional signatures. And the difference between high severity and low severity was not as distinct as it was in the bulk sample transcriptional signature. Despite its limitation, this analysis still offers us more information about where information from genes is mainly used to produce corresponding functional molecular outcomes. If possible, single-cell RNA-seq should be applied later to neutrophils and monocytes for a deeper understanding of how those genes impact the development of the host response to influenza infection at the cellular level.

This study used “limma” (abbreviation for Linear model for Microarray analysis) package to conduct DEA. The model of “limma” assumes that the expression of a gene is the linear combination of the effects of variables. We assumed that there was no interaction between variables. However, this assumption may be incorrect. Other models involving interaction terms can be built to see whether there is a major difference between non-interaction models and with-interaction models.

This study used data after background correction and normalization. However, the technique that the original authors used in the background correction process introduced a lot of negative values and log-transformation of data included a massive loss of information that could have otherwise been preserved [31]. If raw data were made available, a more robust background correction and normalization method combining information from control probes could be used to obtain a more accurate result.

References

1. World Health, O., *Global influenza strategy 2019-2030*. 2019, Geneva: World Health Organization.
2. Longo, D., *HARRISONS PRINCIPLES OF INTERNAL MEDICINE (18th)*. 2011, New York: McGraw-Hill Professional.
3. WHO. *Pandemic (H1N1) 2009 - update 75*. 2009 Nov 2009 [cited 2020 Jan 29]; Available from: https://www.who.int/csr/don/2009_11_20a/en/.
4. Gounder, A.P. and A.C.M. Boon, *Influenza Pathogenesis: The Effect of Host Factors on Severity of Disease.* J Immunol, 2019. **202**(2): p. 341-350.
5. Byrd-Leotis, L., R.D. Cummings, and D.A. Steinhauer, *The Interplay between the Host Receptor and Influenza Virus Hemagglutinin and Neuraminidase.* Int J Mol Sci, 2017. **18**(7).
6. Velkov, T., et al., *The antigenic architecture of the hemagglutinin of influenza H5N1 viruses.* Mol Immunol, 2013. **56**(4): p. 705-19.
7. Reemers, S.S., et al., *Differential gene-expression and host-response profiles against avian influenza virus within the chicken lung due to anatomy and airflow.* J Gen Virol, 2009. **90**(Pt 9): p. 2134-46.
8. Wong, S.S. and R.J. Webby, *Traditional and new influenza vaccines.* Clin Microbiol Rev, 2013. **26**(3): p. 476-92.
9. Heinen, P.P., et al., *Vaccination of pigs with a DNA construct expressing an influenza virus M2-nucleoprotein fusion protein exacerbates disease after challenge with influenza A virus.* J Gen Virol, 2002. **83**(Pt 8): p. 1851-9.
10. Hurt, A.C., *The epidemiology and spread of drug resistant human influenza viruses.* Curr Opin Virol, 2014. **8**: p. 22-9.
11. Mulinari, S. and C. Davis, *Why European and United States drug regulators are not speaking with one voice on anti-influenza drugs: regulatory review methodologies and the importance of 'deep' product reviews.* Health Res Policy Syst, 2017. **15**(1): p. 93.
12. Muthuri, S.G., et al., *Impact of neuraminidase inhibitor treatment on outcomes of public health importance during the 2009-2010 influenza A(H1N1) pandemic: a systematic review and meta-analysis in hospitalized patients.* J Infect Dis, 2013. **207**(4): p. 553-63.
13. Gonzalez-Candelas, F., et al., *Sociodemographic factors and clinical conditions associated to hospitalization in influenza A (H1N1) 2009 virus infected patients in Spain, 2009-2010.* PLoS One, 2012. 7(3): p. e33139.
14. Garcia-Etxebarria, K., et al., *No Major Host Genetic Risk Factor Contributed to A(H1N1)2009 Influenza Severity.* PLoS One, 2015. 10(9): p. e0135983.
15. Everitt, A.R., et al., *IFITM3 restricts the morbidity and mortality associated with influenza.* Nature, 2012. 484(7395): p. 519-23.
16. Park, S.J., et al., *Dynamic changes in host gene expression associated with H5N8 avian influenza virus infection in mice.* Sci Rep, 2015. 5: p. 16512.
17. Tatebe, K., et al., *Response network analysis of differential gene expression in human epithelial lung cells during avian influenza infections.* BMC Bioinformatics, 2010. 11: p. 170.

18. Woods, C.W., et al., *A host transcriptional signature for presymptomatic detection of infection in humans exposed to influenza H1N1 or H3N2.* PLoS One, 2013. **8**(1): p. e52198.
19. Dunning, J., et al., *Progression of whole-blood transcriptional signatures from interferon-induced to neutrophil-associated patterns in severe influenza.* Nature Immunology, 2018. **19**(6): p. 625-635.
20. Klein, E.Y., et al., *The frequency of influenza and bacterial coinfection: a systematic review and meta-analysis.* Influenza Other Respir Viruses, 2016. **10**(5): p. 394-403.
21. Rivas, H.G., S.K. Schmaling, and M.M. Gaglia, *Shutoff of Host Gene Expression in Influenza A Virus and Herpesviruses: Similar Mechanisms and Common Themes.* Viruses, 2016. **8**(4): p. 102.
22. Barbosa-Morais, N.L., et al., *A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data.* 2010. **38**(3): p. e17-e17.
23. Kolde, R., *pheatmap: Pretty Heatmaps*. 2009.
24. Subramanian, A., et al., *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.* Proceedings of the National Academy of Sciences, 2005. **102**(43): p. 15545-15550.
25. Korotkevich, G., V. Sukhov, and A. Sergushichev, *Fast gene set enrichment analysis*. 2016, Cold Spring Harbor Laboratory.
26. Liberzon, A., et al., *Molecular signatures database (MSigDB) 3.0.* Bioinformatics, 2011. **27**(12): p. 1739-1740.
27. Newman, A.M., et al., *Determining cell type abundance and expression from bulk tissues with digital cytometry.* Nat Biotechnol, 2019. **37**(7): p. 773-782.
28. Smyth, G.K. and T. Speed, *Normalization of cDNA microarray data.* 2003. **31**(4): p. 265-273.
29. Dang, C.V., *c-Myc Target Genes Involved in Cell Growth, Apoptosis, and Metabolism.* Molecular and Cellular Biology, 1999. **19**(1): p. 1-11.
30. Toplak, N. and T. Avčin, *Influenza and Autoimmunity.* Annals of the New York Academy of Sciences, 2009. **1173**(1): p. 619-626.
31. Xie, Y., X. Wang, and M. Story, *Statistical methods of background correction for Illumina BeadArray data.* 2009. **25**(6): p. 751-757.
32. Mertz, D., et al., *Populations at risk for severe or complicated influenza illness: systematic review and meta-analysis.* BMJ, 2013. **347**(aug23 1): p. f5061-f5061.
33. Smyth, G.K., *Linear models and empirical bayes methods for assessing differential expression in microarray experiments.* Stat Appl Genet Mol Biol, 2004. **3**: p. Article3.

Figures and tables

Covariate	Statistics	Level	Healthy Control N = 130	Severity I N = 37	Severity II N = 25	Severity III N = 35	Parametric P-value
Age			34.64	38.33	46.20	38.16	**<.001**
Day of illness			0	7.11	7.04	11.11	**<.001**
Sex		Female	75 (57.69)	21 (56.76)	11 (44.00)	20 (57.14)	0.651
		Male	16 (43.24)	16 (43.24)	14 (56.00)	15 (42.86)	
Ethnicity	N (col %)	White	90 (69.23)	25 (67.57)	17 (68.00)	25 (71.43)	0.986
		Other	40 (30.77)	12 (32.43)	8 (32.00)	10 (28.57)	
Pregnancy	N (col %)	Yes	1 (0.77)	17 (45.95)	10 (40.00)	14 (40.00)	**0.003**
		No	74 (56.92)	4 (10.81)	1 (4.00)	6 (17.14)	
		N/A	55 (42.31)	16 (43.24)	14 (56.00)	15 (42.86)	
Comorbidities	N (col %)	0	130 (100)	10 (27.03)	4 (16.00)	13 (37.14)	**<.001**
		1		7 (18.92)	12 (48.00)	14 (40)	
		2		17 (45.95)	6 (24.00)	2 (5.71)	
		≥3		3 (8.11)	3 (12.00)	6 (17.14)	
Bacterial infection status	N (col %)	Yes	130 (100)	20 (54.05)	14 (60.71)	23 (65.71)	**<.001**
		No		17 (45.95)	11 (39.29)	12 (34.29)	

Table 1. Descriptive table of samples grouped by severity.
The parametric p-value is calculated by ANOVA for numerical covariates and chi-square test for categorical covariates.

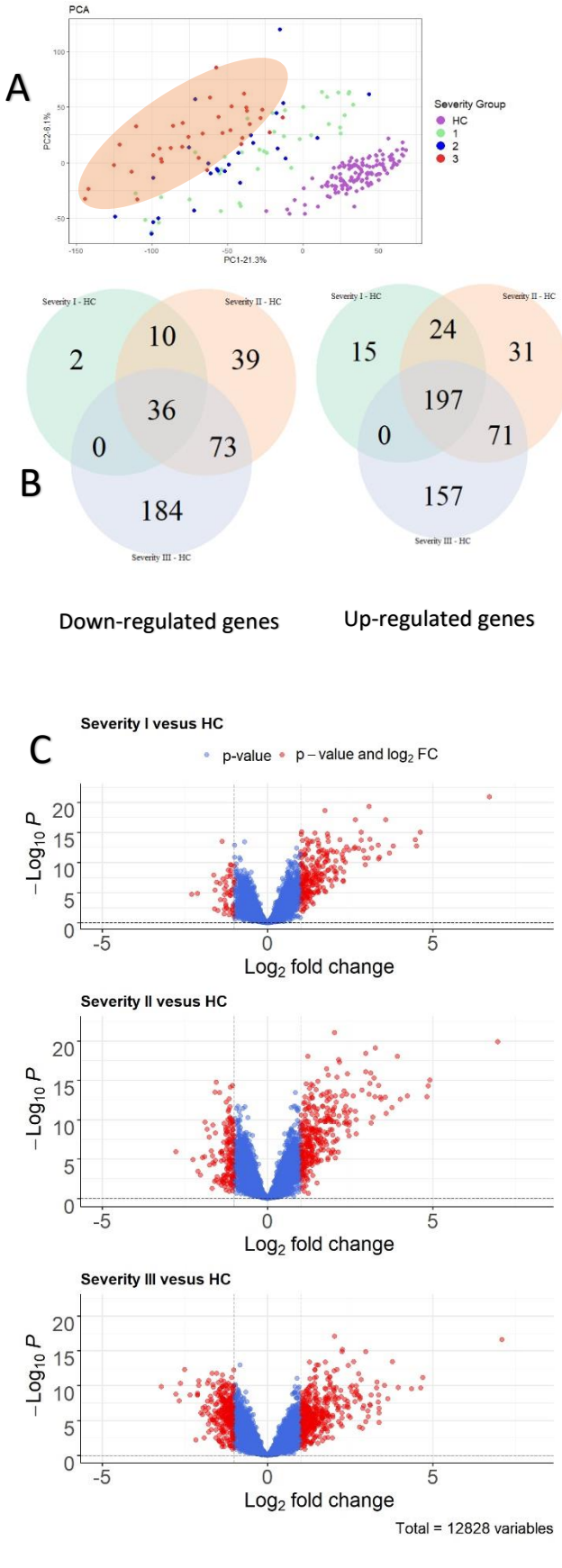
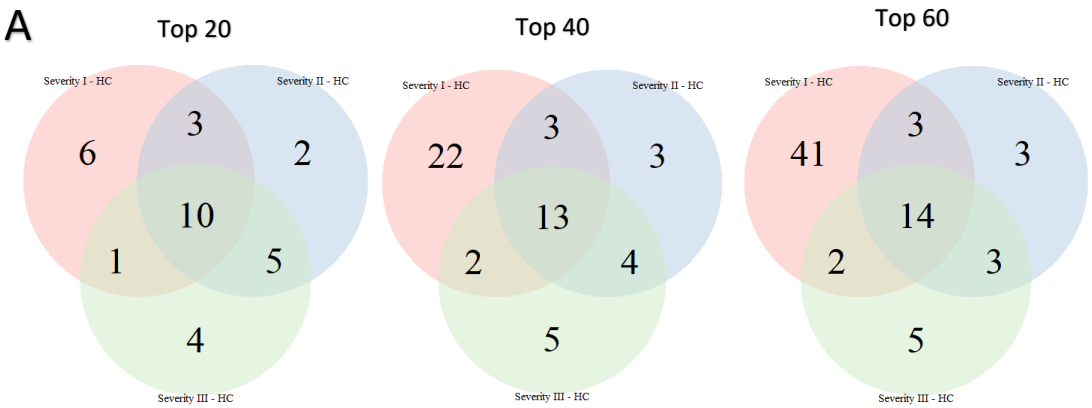

A
PCA
PC2-8.1%
PC1-21.3%
Severity Group
HC
1
2
3
B
Severity I - HC
Severity II - HC
Severity III - HC
2
10
39
36
0
73
184
15
24
31
197
0
71
157
Down-regulated genes
Up-regulated genes
C
Severity I vesus HC
p-value
p – value and log2 FC
−Log10 P
Log2 fold change
Severity II vesus HC
Severity III vesus HC
Total = 12828 variables
D
GZMK
FCER1A
KLRB1
NELL2
CCR3
RNF182
ORM1
RAP1GAP
OSBP2
LOC439949
LEF1
BCL11B
FAIM3
C21orf7
MYOM2
FCRH3
EOMES
CD6
SH2D1A
CD8A
ALDH1A1
FLJ43681
SPOCK2
RASGRP1
HLA-DQB1
ITK
HLA-DMB
MYBL1
SKAP1
IL2RB
S1PR5
TGFBI
THEMIS
CLIC3
CD247
CD160
HLA-DPA1
PVRIG
SAMD3
HLA-DQA1
ZFP36L2
PRKCH
KLRF1
CSF1R
GPR56
FGFBP2
TGFBR3
GZMA
GZMH
KLRG1
PRF1
HLA-DRB6
GIMAP5
PLEKHA1
IL7R
PYHIN1
SBK1
CPVL
ANKRD22
C19orf59
IL18R1
UGCG
CLEC4D
RNASE2
PRTN3
LCN2
RNASE3
TCN1
CTSG
CEACAM8
CEACAM6
MPO
DEFA4
ELANE
RETN
MMP8
DEFA1
HP
ARG1
TDRD9
GPR84
RGL4
MMP9
OPLAH
CD177
ECRP
CST7
CKAP4
ANXA3
GYG1
UPP1
ADCY3
CEACAM1
NLRC4
SRPK1
PGD
PFKFB2
DYSF
OSTalpha
FOLR3
C5orf32
CLEC5A
S100A12
UPB1
BMX
TOP2A
SCD
ATP8B4
CAMP
LOC284648
PTPN20A
CEBPE
TACSTD2
BEX1
STOM
LOC284757
SAMD14
PCOLCE2
ZDHHC19
COL17A1
MAOA
ABCA13
AZU1
OLFM4
LTF
BPI
OLR1
LOC653061
SLC2A5
MS4A3
CD24
PGLYRP1
HIST1H4H
SLFN5
CD3G
KLRD1
GNLY
CX3CR1
severity1
severity2
severity3

Figure 1. Gene-level analysis.

(A)Principal component analysis. *X-axis: the highest principal component accounting for the variance. Y-axis: the second-highest principal component accounting for the variance. Samples were annotated by different colors by their severity status: healthy control (HC, purple), severity I (green), severity II (blue), and severity III (red)* ***(B) Overlaps of DEGs in three contrasts.*** *The Venn diagram showed the pairwise and three-way overlap between differentially expressed genes (**DEGs**) in Severity I – HC, Severity II – HC, and Severity III – HC contrasts. Genes in pairwise overlap are shared by both contrasts. The overlap of pairwise overlaps determines the three-way overlap. Left panel: genes were down-regulated; Right panel: genes were up-regulated.* ***(C) Volcano plots of differentially expressed genes under different severities compared with healthy control.*** *Genes had log2-fold change larger than 1 and adjusted p-value less than 0.1 (considered as DEGs in this study) were annotated as red dots. Genes had log2-fold change larger than 1 but adjusted p-value did not meet 0.1 criteria were annotated as blue dots.* ***(D) Heatmap of 138 DEGs that were differently regulated among different severity stages compared with healthy controls.*** *Genes were selected from distinct 824 DEGs in DEA with the following criteria: there is a difference of at least 1 in log2-fold-change between any two of the three contrasts (Severity I – HC, Severity I – HC, and Severity III – HC). The color showed the log2-fold-change of the 138 DEGs under the three contrasts.*


A
Top 20
Severity I - HC
Severity II - HC
Severity III - HC
6
3
2
10
1
5
4
Top 40
Severity I - HC
Severity II - HC
Severity III - HC
22
3
3
13
2
4
5
Top 60
Severity I - HC
Severity II - HC
Severity III - HC
41
3
3
14
2
3
5

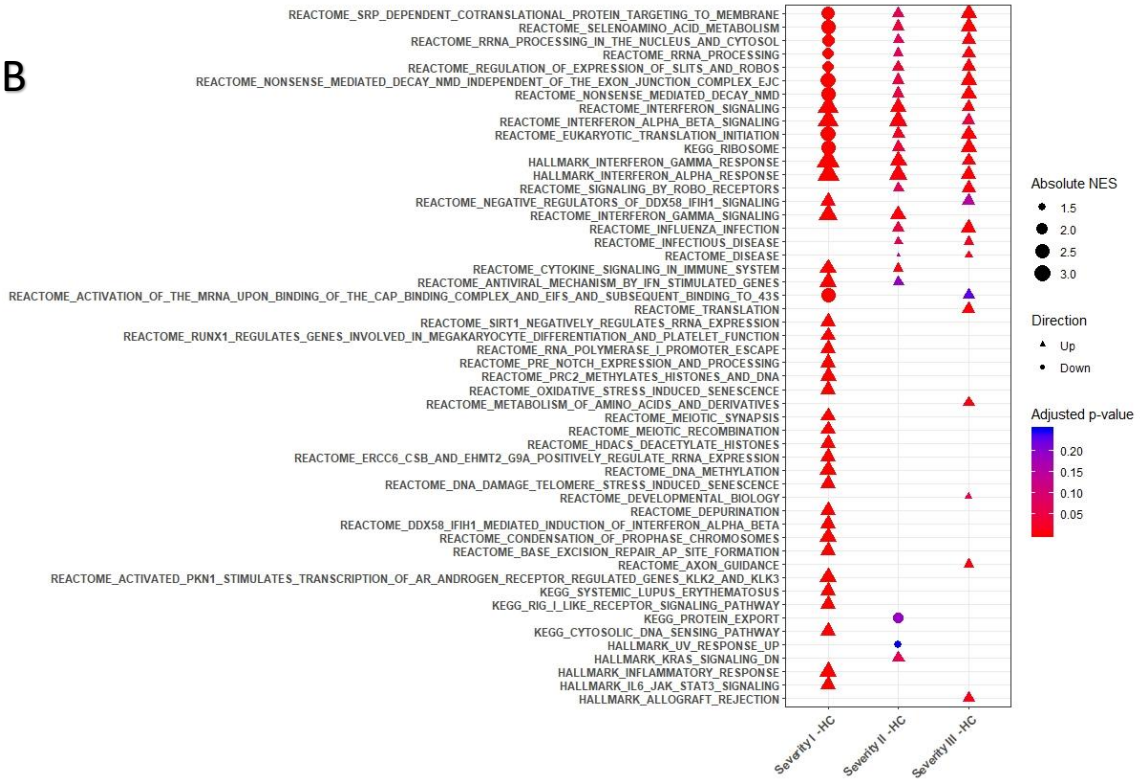

B
REACTOME_SRP_DEPENDENT_COTRANSLATIONAL_PROTEIN_TARGETING_TO_MEMBRANE
REACTOME_SELENOAMINO_ACID_METABOLISM
REACTOME_RRNA_PROCESSING_IN_THE_NUCLEUS_AND_CYTOSOL
REACTOME_RRNA_PROCESSING
REACTOME_REGULATION_OF_EXPRESSION_OF_SLITS_AND_ROBOS
REACTOME_NONSENSE_MEDIATED_DECAY_NMD_INDEPENDENT_OF_THE_EXON_JUNCTION_COMPLEX_EJC
REACTOME_NONSENSE_MEDIATED_DECAY_NMD
REACTOME_INTERFERON_SIGNALING
REACTOME_INTERFERON_ALPHA_BETA_SIGNALING
REACTOME_EUKARYOTIC_TRANSLATION_INITIATION
KEGG_RIBOSOME
HALLMARK_INTERFERON_GAMMA_RESPONSE
HALLMARK_INTERFERON_ALPHA_RESPONSE
REACTOME_SIGNALING_BY_ROBO_RECEPTORS
REACTOME_NEGATIVE_REGULATORS_OF_DDX58_IFIH1_SIGNALING
REACTOME_INTERFERON_GAMMA_SIGNALING
REACTOME_INFLUENZA_INFECTION
REACTOME_INFECTIOUS_DISEASE
REACTOME_DISEASE
REACTOME_CYTOKINE_SIGNALING_IN_IMMUNE_SYSTEM
REACTOME_ANTIVIRAL_MECHANISM_BY_IFN_STIMULATED_GENES
REACTOME_ACTIVATION_OF_THE_MRNA_UPON_BINDING_OF_THE_CAP_BINDING_COMPLEX_AND_EIFS_AND_SUBSEQUENT_BINDING_TO_43S
REACTOME_TRANSLATION
REACTOME_SIRT1_NEGATIVELY_REGULATES_RRNA_EXPRESSION
REACTOME_RUNX1_REGULATES_GENES_INVOLVED_IN_MEGAKARYOCYTE_DIFFERENTIATION_AND_PLATELET_FUNCTION
REACTOME_RNA_POLYMERASE_I_PROMOTER_ESCAPE
REACTOME_PRE_NOTCH_EXPRESSION_AND_PROCESSING
REACTOME_PRC2_METHYLATES_HISTONES_AND_DNA
REACTOME_OXIDATIVE_STRESS_INDUCED_SENESCENCE
REACTOME_METABOLISM_OF_AMINO_ACIDS_AND_DERIVATIVES
REACTOME_MEIOTIC_SYNAPSIS
REACTOME_MEIOTIC_RECOMBINATION
REACTOME_HDACS_DEACETYLATE_HISTONES
REACTOME_ERCC6_CSB_AND_EHMT2_G9A_POSITIVELY_REGULATE_RRNA_EXPRESSION
REACTOME_DNA_METHYLATION
REACTOME_DNA_DAMAGE_TELOMERE_STRESS_INDUCED_SENESCENCE
REACTOME_DEVELOPMENTAL_BIOLOGY
REACTOME_DEPURINATION
REACTOME_DDX58_IFIH1_MEDIATED_INDUCTION_OF_INTERFERON_ALPHA_BETA
REACTOME_CONDENSATION_OF_PROPHASE_CHROMOSOMES
REACTOME_BASE_EXCISION_REPAIR_AP_SITE_FORMATION
REACTOME_AXON_GUIDANCE
REACTOME_ACTIVATED_PKN1_STIMULATES_TRANSCRIPTION_OF_AR_ANDROGEN_RECEPTOR_REGULATED_GENES_KLK2_AND_KLK3
KEGG_SYSTEMIC_LUPUS_ERYTHEMATOSUS
KEGG_RIG_I_LIKE_RECEPTOR_SIGNALING_PATHWAY
KEGG_PROTEIN_EXPORT
KEGG_CYTOSOLIC_DNA_SENSING_PATHWAY
HALLMARK_UV_RESPONSE_UP
HALLMARK_KRAS_SIGNALING_DN
HALLMARK_INFLAMMATORY_RESPONSE
HALLMARK_IL6_JAK_STAT3_SIGNALING
HALLMARK_ALLOGRAFT_REJECTION
Severity I -HC
Severity II -HC
Severity III -HC
Absolute NES
1.5
2.0
2.5
3.0
Direction
Up
Down
Adjusted p-value
0.20
0.15
0.10
0.05

Figure 2. Gene set enrichment analysis.

(A) Overlaps of enriched gene sets. *This Venn diagram showed the pairwise and three-way overlap between enriched gene sets in Severity I – HC, Severity II – HC, and Severity III – HC contrasts. Gene sets in pairwise overlap are shared by both contrasts. The overlap of pairwise overlaps determines the three-way overlap. Left panel: top 20 gene sets by absolute NES after filtered by FDR < 0.25; Middle panel: top 40 gene sets by absolute NES after filtered by FDR adjusted p-value < 0.25; Right panel: top 60 gene sets by absolute NES after filtered by FDR adjusted p-value < 0.25.* ***(B) Enrichment plot of the top 40 gene sets from each of the three contrasts ordered by absolute Normalized Enrichment Score (NES).*** *Gene Set Enrichment Analyses (GESA) was conducted using all genes mapped from the microarray and their fold-change information were used for ranking. Four representative information: Normalized Enrichment Score (NES), Adjusted P-value, Direction of regulation were shown for 52 distinct gene sets and 3 contrasts.*

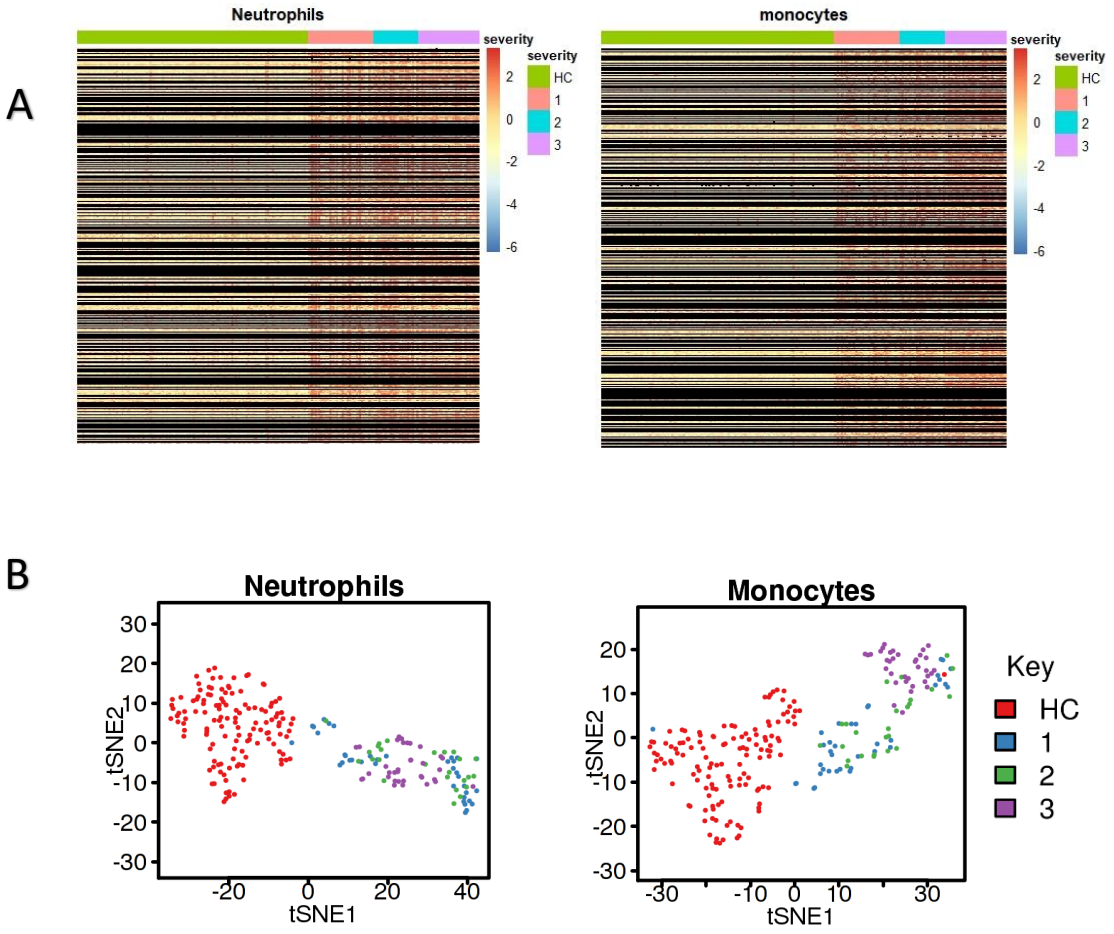


Figure 3. Cell-type-specific analysis.

(A) Heatmaps of cell-specific transcriptional profiles. *Left panel: neutrophils; Right panel: monocytes. A high-resolution expression profile (HREP) of 227 samples was available for each of the 10 major LM22 cell types. In each HREP, only the expressions of 824 DEGs were available. Data were log2 transformed and mean-centered before plotting.* ***(B) Two-dimensional t-SNE plots profiled from 227 samples by CibersortX***. *t-SNE plots were generated by HREPs for each of the 10 LM22 major cell types (only two of them were shown here). Each dot in the plot represents the expressions of one of the 227 samples in corresponding cell types. Each sample was color-coded according to severity.*

Appendix 1: Supplementary figure and table

Covariate	Statistics	Level	Influenza N = 61	Healthy control N = 130	Parametric P-value
Age	Mean (SD)		38.98 (11.98)	34.64 (11.12)	**0.015**
Sex	N (col %)	Female	33 (54.10)	21 (56.76)	0.640
		Male	28 (45.90)	16 (43.24)	
Ethnicity	N (col %)	White	43 (70.49)	90 (69.23)	0.860
		Other	18 (29.51)	40 (30.77)	
Pregnancy	N (col %)	Yes	7 (11.48)	1 (0.77)	**0.001**
		No	26 (42.62)	74 (56.92)	
		N/A	28 (45.90)	55 (42.31)	
Comorbidities	N (col %)	0	17 (27.87)	130 (100)	**0.004**
		1	20 (32.79)		
		2	17 (27.87)		
		≥3	7 (11.48)		
Bacterial infection status	N (col %)	Yes	35 (57.38)		**<.001**
		No	26 (42.62)	130 (100)	

Supplementary Table 1.

Descriptive table of subjects.

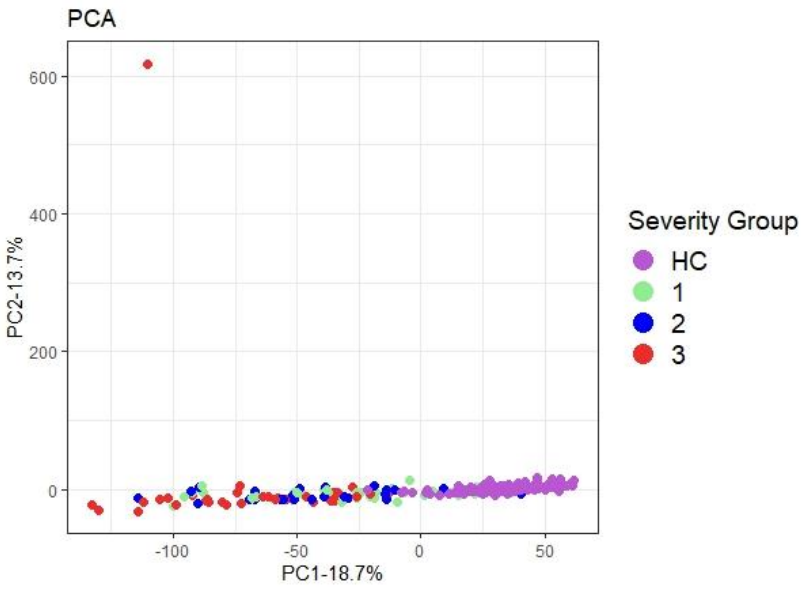
The parametric p-value is calculated by ANOVA for numerical covariates and chi-square test for categorical covariates.

Gene Set Name	# Genes in Gene Set (K)	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
REACTOME_NEUTROPHIL_DEGRANULATION	478	Neutrophil degranulation	41	0.0858	7.26E-45	1.47E-41
REACTOME_INNATE_IMMUNE_SYSTEM	1104	Innate Immune System	49	0.0444	4.19E-40	4.24E-37
REACTOME_ANTIMICROBIAL_PEPTIDES	97	Antimicrobial peptides	12	0.1237	1.46E-15	9.82E-13
HALLMARK_ALLOGRAFT_REJECTION	200	Genes up-regulated during transplant rejection.	12	0.06	9.14E-12	4.63E-09
KEGG_ASTHMA	30	Asthma	6	0.2	1.03E-09	4.18E-07
REACTOME_IMMUNOREGULATORY_INTERACTIONS_BETWEEN_A_LYMPHOID_AND_A_NON_LYMPHOID_CELL	186	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	10	0.0538	1.52E-09	5.14E-07
REACTOME_ADAPTIVE_IMMUNE_SYSTEM	811	Adaptive Immune System	17	0.021	5.68E-09	1.64E-06
KEGG_GRAFT_VERSUS_HOST_DISEASE	41	Graft-versus-host disease	6	0.1463	7.56E-09	1.91E-06
KEGG_ALLOGRAFT_REJECTION	37	Allograft rejection	5	0.1351	2.15E-07	4.36E-05
REACTOME_GENERATION_OF_SECOND_MESSENGER_MOLECULES	37	Generation of second messenger molecules	5	0.1351	2.15E-07	4.36E-05

Supplementary Table 2.

Important gene sets in over-representative analysis.

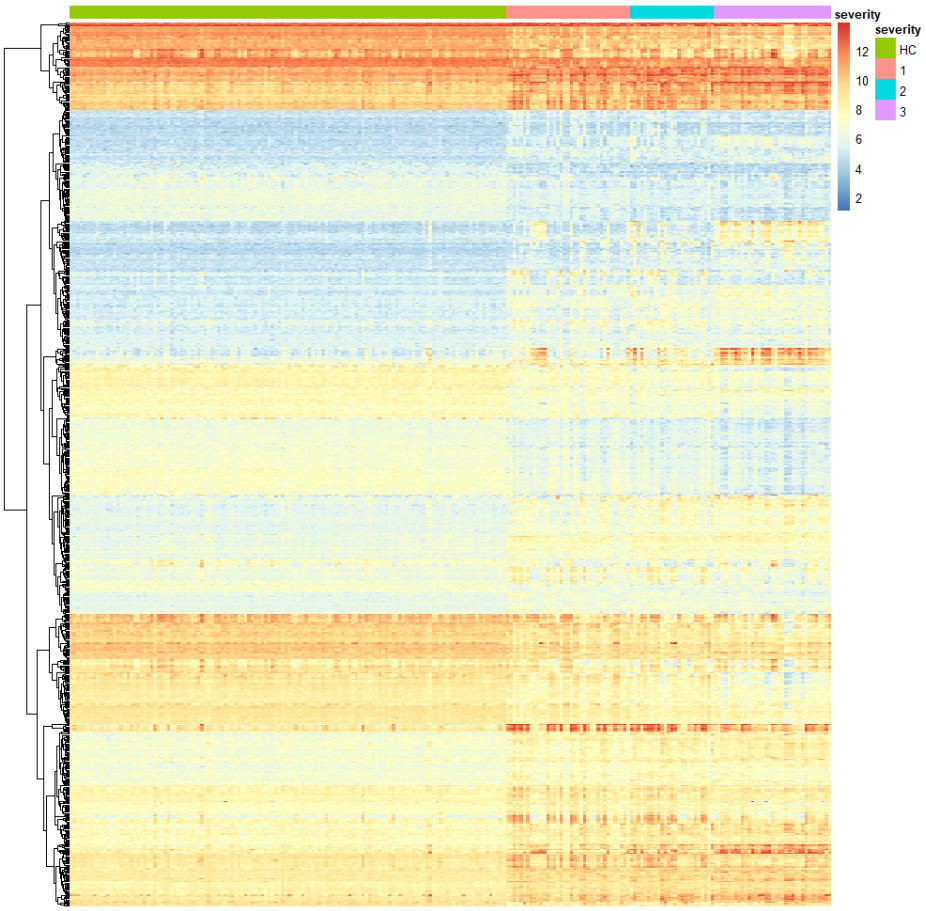
The top 10 gene sets after filtered by FDR q-value < 0.05



Supplementary Figure 1.

Plot of the first two principal components before removing the outlier.

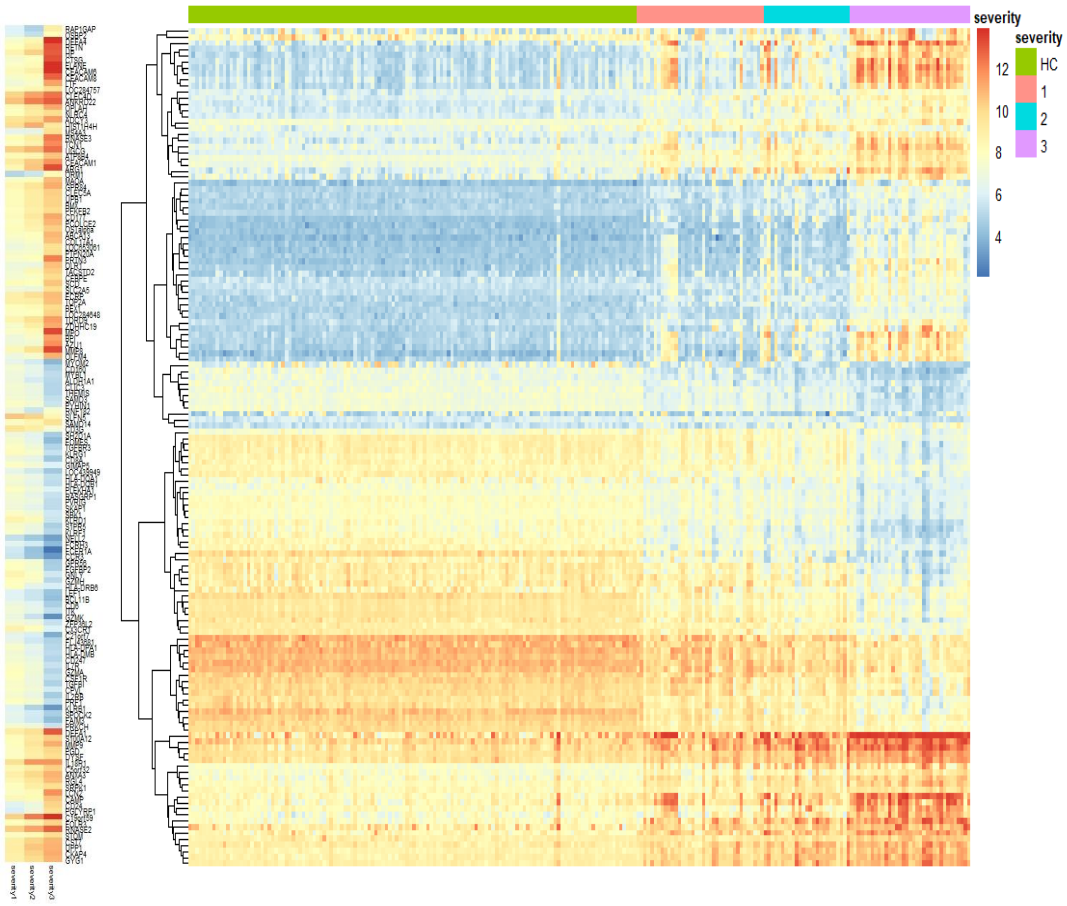
A major of variance in the PC2 was caused by the outlier making other samples almost lie in a line in the PC2.



Supplementary Figure 2.

Heatmap of 824 distinct differentially expressed genes.

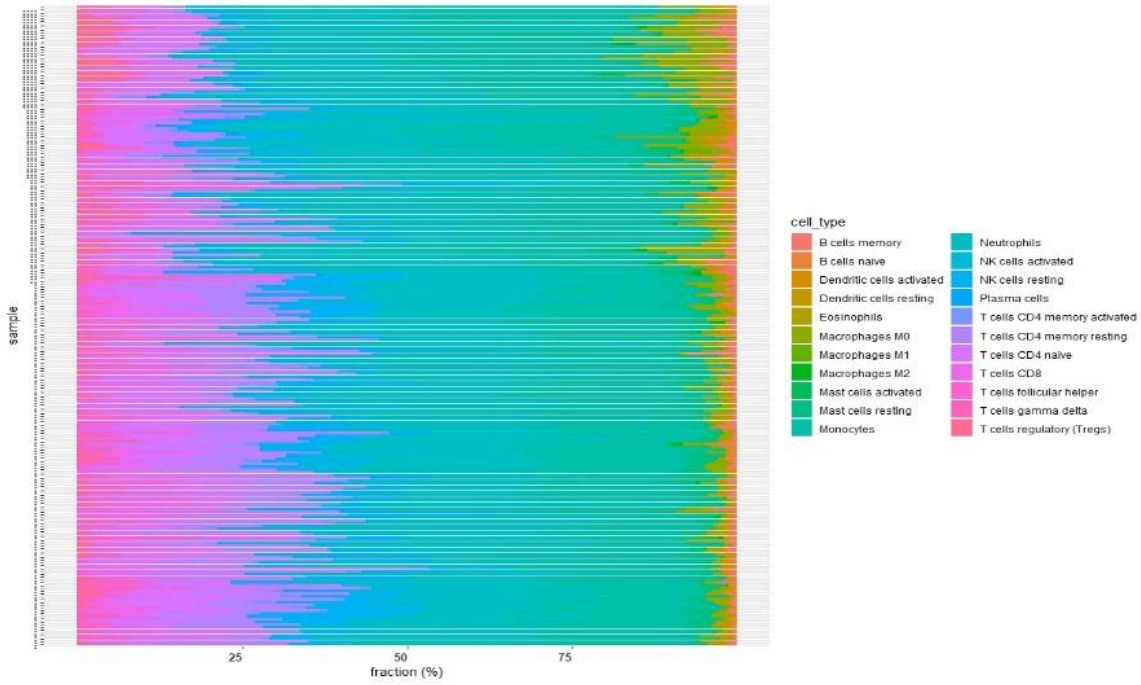
The figure shows different transcription signatures between the healthy control group and the three influenza A severity groups.



Supplementary Figure 3.

Heatmap of 138 DEGs that were differently regulated among different severity stages compared with healthy controls.

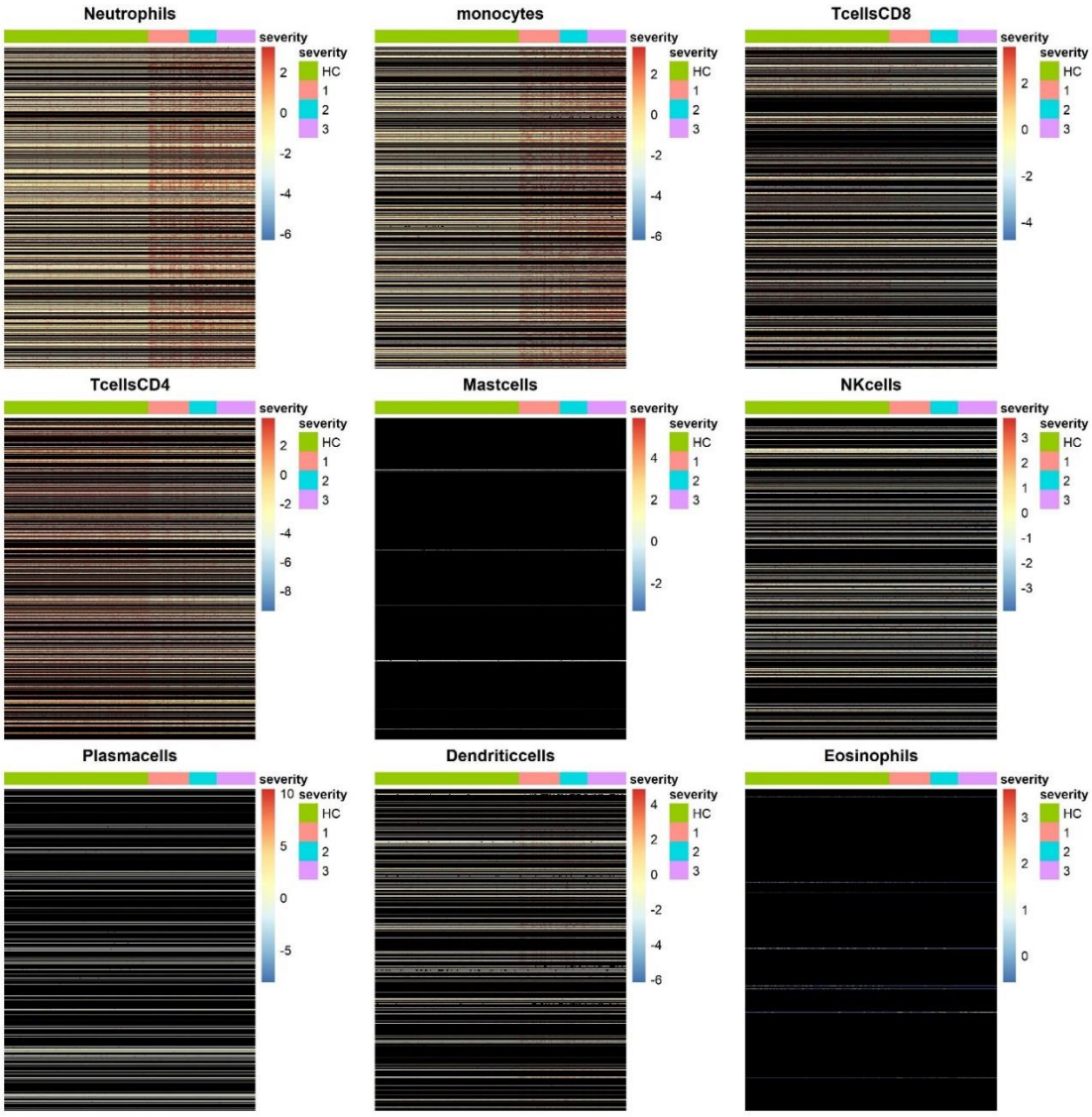
Genes were selected from distinct 824 DEGs in DEA with the following criteria: at least one of the three pairwise comparisons of fold-changes from three contrasts (Severity I – HC, Severity I – HC, and Severity III – HC) show large enough difference (there is a difference of at least 1 in log2-fold-change between any two of the three contrast). Left panel: the fold-changes of the 138 DEGs under three contrasts (Severity I – HC, Severity I – HC, and Severity III – HC). Right panel: the normalized intensities of the 138 DEGs under different conditions (HC, Severity I, Severity II, and Severity III).



Supplementary Figure 4.

Fractions of 22 cell-types in all samples.

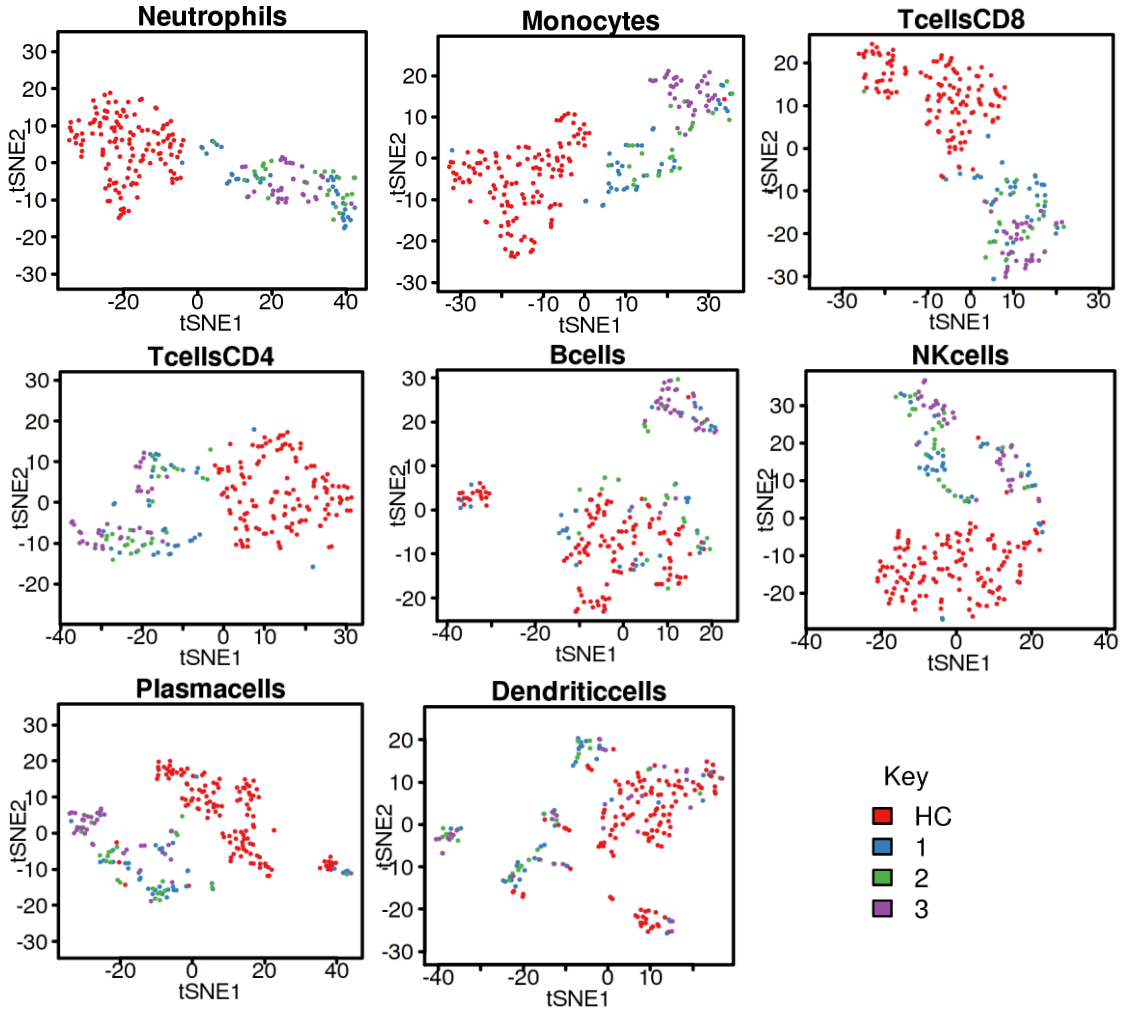
Left panel: Cell-type fractions were calculated by the CibersortX and represented as percentage amounted to 100% on the x-axis. 227 samples were arranged on the y-axis. The number of dots on the y-axis represents the severity of that sample: "." for healthy controls, ".." for severity I group, "..." for severity II group and "...." for severity III group. Each cell type was represented by a different color and similar cell types were represented as similar colors. Right panel: names and color-codes of corresponding 22 cell-types.



Supplementary Fig 5.

Heatmaps of cell-specific transcriptional profiles.

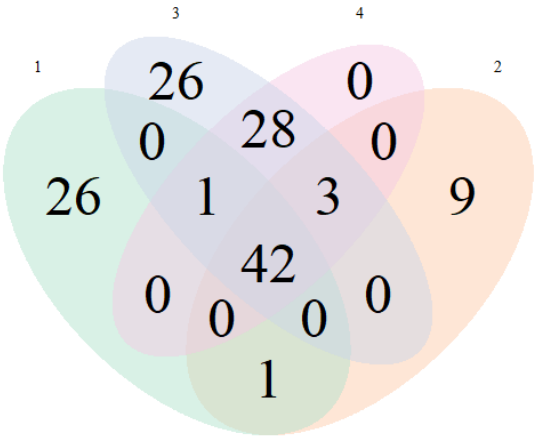
High-resolution expression profiles (HREPs) of 227 samples were available for each of the 10 major LM22 cell types. In each HREP, only the expressions of 824 DEGs were available. Data were log2 transformed and mean-centered before plotting. Genes that didn't have information in imputed cell-specific transcriptional profiles were denoted as black.



Supplementary Figure 6.

Two-dimensional t-SNE plots profiled from 227 samples by CibersortX.

t-SNE plots were generated by HREPs for each of the 10 LM22 major cell types. Each dot in the plot represents the expressions of one of the 227 samples in corresponding cell types. The bottom-right corner shows that each sample was color-coded by its severity.



Supplementary Figure 7.

Venn diagram of the four universally enriched gene set in high severity stage in the REACTOME collection.

1: REACTOME_TRANSLATION,

2: REACTOME_METABOLISM_OF_AMINO_ACIDS_AND_DERIVATIVES,

3: REACTOME_DEVELOPMENTAL_BIOLOGY,

4: REACTOME_AXON_GUIDANCE

Appendix 2: Supplementary methods

Design matrix for differentially expression analysis (DEA)

A design matrix was built based on a series of covariates: age, pregnancy status, days of illness, comorbidities, bacteria status, and severity. Among those covariates, only days of illness was treated as a continuous variable. Since elderly people which commonly defined as older than 65 years are more likely to have severe influenza [32], we use this rule to make age as a binary variable in this study. Besides, pregnancy status (male with N/A was coded as No) and bacterial status were also treated as binary variables. Comorbidities (0, 1, 2, and ≥3), and severity (HC, 1, 2, and 3) were multi-level factors.

Bayesian hierarch model for this study

$$E(y_{gi}) = \beta_{g0} + \beta_{g1} * (Severity\ I)_i + \beta_{g2} * (Severity\ II)_i + \beta_{g3} * (Severity\ III)_i + \beta_{g4} * (Bacteria\ infection\ status)_i + \beta_{g5} * (Pregency\ status)_i + \beta_{g6} * (Binary\ age)_i + \beta_{g7} * (comorbidities: no) + \beta_{g8} * (comorbidities: 1) + \beta_{g9} * (comorbidities: 2) + \beta_{g10} * (comorbidities: \geq 3) + \beta_{g11} * (Day\ of\ illness)$$

β_{g0}: the basic line of gene g expression in log2 scale. The basic line condition is a healthy control group, no bacteria infection, no pregnancy, age ≤ 65, no comorbidity.

β_{gj}: the effect of jth factor in gene g's expression in log2 scale, $j = 1, 2, \ldots 10$.

$\beta_{g1},\ \beta_{g2}, and\ \beta_{g3}$ will be the three contrasts in the method section respectively since *the* intercept β_0 will act as the basic line.

Y_{gi}: log2 normalized intensity of gene g in sample i.

Let set vector $\boldsymbol{y}_g$ as the expression of gene g of all the sample and β_g as the coefficient vector.

Then the above equation can be simplified as

$$E(\boldsymbol{y}_g) = \boldsymbol{X} \cdot \boldsymbol{\beta_g}$$

$X\ is\ the\ design\ matrix\ with\ each\ row\ as\ a\ sample\ and\ each\ column\ as$ a character of that sample

The above linear model is fitted to the log2 intensity of each gene to get corresponding $\hat{\alpha}_g,\ esitmate\ of\ \sigma_g^2$: $s_g^2, and\ estimated\ covariance\ matrix$:

$$\widehat{var}(\hat{\beta}_g) = V_g\, S_g^2$$

$$V_g\ is\ a\ positive\ definite\ matrix\ independent\ of\ s_g^2$$

Let annotate ν_{gj} as the jth diagonal element of V_g . Assume that:

$$\hat{\beta}_{gj} \mid \beta_{gj}, \sigma_g^2 \sim N(\alpha_{gj}, v_{gj}\sigma_g^2)$$

And

$$s_g^2 \mid \sigma_g^2 \sim \frac{\sigma_g^2}{d_g}\, \chi^2_{d_g}$$

$d_g$is the residual degree of freedom for the linear model of gene g

Assume that the gene specific parameters such as $\sigma_g^{2'} s\ actually\ dependent$ with each other.

We construct such a correlation structure by using a bayesian hierarchical model and set prior with some common hyperparameter shared by different genes:

The prior of σ_g^2 $is\ a\ inverse\ scaled\ chisquare\ distribution$:

$$\frac{1}{\sigma_g^2} \sim \frac{1}{d_0 s_0^2}\ \chi_{d_0}^2,$$

which is a conjugate prior for normal distribution with known mean μ

The posterior distribution of σ_g^2 is an inverse scaled chi − square with $d_0 + d_g$ as the degree parameter and $\frac{d_0 S_0^2 + d_g S_g^2}{d_0 + d_g}$ as the scale parameter.

Besides the posterior mean denoted $as\ \tilde{s}_g$ is the scale parameter:

$$\tilde{s}_g = \frac{d_0 S_0^2 + d_g S_g^2}{d_0 + d_g}$$

$The\ prior\ for\ \beta_{gj}\ is\ set\ as$:

$$\beta_{gj} \mid \sigma_g^2 \sim N\left(0, v_{0j}\sigma_g^2\right)$$

It has been shown that:

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{v_{gj}}}$$

follow a t-distribution with degrees of freedom $d_g + d_0$.

If we just don't impose a hierarchical structure to obtain a correlation between different genes, we can also obtain a t-statistic assuming $\hat{\beta}_{gj}\ and\ s_g^2\ are\ independent$

$$t_{gj} = \frac{\hat{\beta}_{gj}}{s_g \sqrt{v_{gj}}} \qquad t_{gj} \text{ follows a t} - \text{distribution with degrees of freedom } d_g$$

The distribution of $\tilde{t}_{gj}$ has d_0 more degrees of freedom than t_{gj}. Since t_{gj} is a statistic under frequentist standard, and t_{gj} is a hybrid of bayesian and frequentist, we call it moderated t-statistic. The extra degrees of freedom reflect the information borrowed from other genes.

The hyperparameters $v_{0j}, s_0,$ and d_0 are decided by the empirical Bayes method implemented in the "limma" package [33].