Predicting seasonal influenza vaccine response using systemic gene expression profiling

Christian V. Forst[1,6], Matthew Chung[2], Megan Hockman[3], Lauren Lashua[3], Emily Adney[3], Michael Carlock[4,5], Ted Ross[4,5], Elodie Ghedin[2], and David Gresham[3,6]

[1]Genetics and Genomic Sciences, Department of Microbiology, Icahn School of Medicine at Mt Sinai
[2]Systems Genomics Section, Laboratory of Parasitic Diseases, NIAID, NIH, Bethesda, MD 20894, USA
[3]Center for Genomics and Systems Biology, Department of Biology, New York University, New York, NY 10003, USA
[4]Center for Vaccines and Immunology, University of Georgia, Athens, GA 30602, USA
[5]Department of Infectious Diseases, University of Georgia, Athens, Georgia 30602, USA
[6]Correspondence: <christian.forst@mssm.edu > and <dgresham@nyu.edu>

Abstract

Seasonal influenza is a primary public health burden in the USA and globally. Annual vaccination programs are designed on the basis of circulating influenza viral strains. However, the effectiveness of the seasonal influenza vaccine is highly variable between seasons and among individuals. A number of factors are known to influence vaccination effectiveness including age, sex, and comorbidities. Here, we sought to determine whether whole blood gene expression profiling prior to vaccination is informative about pre-existing immunological status and the immunological response to vaccine. We performed whole transcriptome analysis using RNA sequencing (RNAseq) of whole blood samples obtained prior to vaccination from participants enrolled in an annual influenza vaccine trial. Serological status prior to vaccination and 28 days following vaccination was assessed using the hemagglutination inhibition assay (HAI) to define baseline immune status and the response to vaccination. We find evidence that genes with immunological functions are increased in expression in individuals with higher pre-existing immunity and in those individuals who mount a greater response to vaccination. Using a random forest model we find that this set of genes can be used to predict vaccine response with a performance similar to a model that incorporates physiological and prior vaccination status alone. Our study shows that increased expression of immunological genes, possibly reflecting greater plasmablast cell populations, prior to vaccination is associated with an enhanced response to vaccine. Furthermore, in the absence of physiological information and vaccination history, whole blood gene expression signatures are informative about the immunological response of an individual to seasonal influenza vaccination.

Introduction

In non-pandemic years, seasonal influenza generally imposes considerable mortality and morbidity burdens resulting in three to five million severe illnesses per year and 290,000 to 650,000 respiratory deaths worldwide (WHO 2018). Public health initiatives to annually administer the seasonal influenza vaccine, targeting predominant circulating influenza strains, are an effective means of mitigating the impact of the virus. However, the effectiveness of the seasonal vaccine - defined as the percent reduction in the frequency of influenza illness among vaccinated individuals compared to the frequency among unvaccinated individuals (CDC 2021b) - shows considerable variation between seasons and among individuals within a season. Between season variation in vaccine effectiveness is primarily attributed to mismatches between the vaccine strain selection and circulating viruses. By contrast, within season variation in vaccine effectiveness is impacted by both viral and host factors, including age, preexisting health conditions such as obesity (Neidich et al. 2017), and prior vaccination and infection history. Understanding the sources of interindividual variation in vaccine effectiveness is critical to the development of robust vaccination strategies.

Host biomolecular factors may be an underlying source of interindividual variation in vaccine effectiveness, specifically antibody repertoires. A number of observational studies provide evidence that pre-existing immunity, due to either prior infection or vaccination, may impact the immune response to vaccination, a phenomenon known as “original antigenic sin” (Cobey and Hensley 2017, Jang and Ross 2019). The immune response to infection may be dominated by antibodies to previously-encountered hemagglutinin (HA) epitopes, resulting in varying vaccine effectiveness that depends largely on an individual’s age and vaccination history. Additionally, antigenic similarity between consecutive vaccine strains may dampen an individual’s ability to generate antibodies following repeated vaccinations. Antibodies from the initial encounter may mask similar vaccine epitopes in the second encounter, resulting in a diminished response, a phenomenon known as the antigenic distance hypothesis (Smith et. al. 1999, Skowronski et. al. 2016).

Preexisting immunity and vaccine response can be assessed using the hemagglutination inhibition (HAI) assay, which detects the presence of antibodies that prevent the HA protein of the influenza virus from agglutinating red blood cells (Katz, Hancock, and Xu 2011). Vaccine response is quantified on the basis of an increase in antibody titer, which can be inferred using the HAI assay for which a 4-fold rise in antibody titers is typically used to define seroconversion (Katz, Xu, 2011; Cauchemez, 2012). Although not a direct measure of vaccine effectiveness, there is strong evidence that seroconversion is predictive of vaccine effectiveness (Cowling et al. 2019). However, a caveat to the serological-based assessment of vaccine response using the HAI assay is that it may be inaccurate if most antibodies target the neuraminidase (Chen et al. 2018).

Prior studies have shown that there is considerable variation in the capacity to respond to vaccination (Stacey et al. 2018) with reduced responsiveness associated with biological variables such as sex (Klein and Pekosz 2014). Differences in gene expression may be related to, and predictive of, variation in vaccine effectiveness. For example, one large-scale study identified nine genes and three gene modules that show significant expression variation associated with the magnitude of the antibody response (HIPC-CHI Signatures Project Team and HIPC-I Consortium 2017). Other studies have identified age-associated differences in gene expression that are associated with differences in vaccine response (Thakar et al. 2015). These studies indicate that gene expression profiles could be informative about how well an individual will respond to vaccination. However, whether these findings hold in other cohorts and the interactive effects of biological and molecular factors on the predictive value of gene expression in the context of the vaccine response have not previously been assessed.

In this study we sought to determine whether whole blood mRNA expression profiles could be used to help predict the response to influenza vaccination. We analyzed gene expression prior to vaccination and identified genes that are differentially expressed reflecting differing pre-vaccination immune states. We then tested whether gene expression differences were associated with differences in vaccine response among individuals. We find that a predictive model of vaccine response that uses only the gene expression state of individuals prior to vaccination shows comparable performance to a model that uses only physiological factors (e.g. BMI, sex, and age) and knowledge of prior vaccination history. However, a model that incorporates both gene expression and physiological and demographic factors does not outperform either of these models indicating that they do not contain additive information. Our study demonstrates that gene expression states prior to vaccination show variation that reflects the preexisting immunological state of individuals and these profiles have predictive value with respect to vaccine response that may be of use when physiological information and vaccination history is not available.

Results

To investigate the utility of global gene expression profiling in predicting the response to seasonal influenza vaccination, we studied a cohort of participants recruited by the University of Georgia (UGA). The study included a total of 275 participants from a cohort herein referred to as UGA4 that was conducted during the 2019-2020 influenza season. Volunteers provided a blood sample for baseline serology prior to vaccination (Day 0). They were then given a commercially available, seasonal influenza split inactivated virus vaccine (Fluzone, Sanofi Pasteur), and subsequently provided a blood sample 28 days after vaccination to assess response serology. The vaccine was formulated with influenza A strains for H1N1 (A/Brisbane/2/2018) and H3N2 (A/Kansas/14/2017), in addition to influenza B strains for Victoria (B/Colorado/6/2017) and Yamagata (B/Phuket/3073/2013). Standard dose (SD) quadrivalent vaccines (0.5mL dose with 15μg HA/strain or 60μg HA total) were administered to most participants (n = 214); however, those participants aged 65 years and older were given the option of a high-dose (HD) formulation (trivalent 0.5mL dose with 60μg HA/strain or 180μg HA total; n = 61) in which the Yamagata component was excluded.

The 275 study participants included adult males (n = 102) and females (n = 173), ranging in age from 18-85 years (mean ± one standard deviation age = 50.3 ± 17.2 years). Self-declared race was recorded: study participants predominantly self-identified as white (n=220), with a small number of black (n = 27), hispanic (n = 17), asian (n= 6), and Native American (n = 2) participants. Body mass index (BMI) of participants ranged from 17.43 to 59.1 with 199 of the 275 study participants exceeding a BMI of 25 (**Table 1**) which is classified as overweight using standard CDC definitions (CDC 2021a). Participants were assessed for risk factors including smoking (n = 94 previous or current smokers) and comorbidities of which hypertension was the most common (n = 49). Overall, health metrics are consistent with enrichment for high risk factors among study participants.

Table 1. Demographic and health metrics of study participants. A total of 275 participants in the UGA4 (2019-2020) split inactivated virus (Fluzone, Sanofi Pasteur) influenza vaccine trial were included in the study.

Metric	Categories	Mean (+/- standard deviation)	Range
Age		50.3 ± 17.2 years)	18 - 85 years
Sex	Males 102 Female 173		-
Race/Ethnicity	White 220 Black 27 Hispanic 17 Asian 6 Native American 2 Multiracial 3	-	-
BMI	Overweight (BMI ≥ 25) 199 Healthy (BMI 18.5 - 25) 76	29.36 (± 6.66)	17.43 - 59.1
Smoking	Previously 73 Current 21 Never 179 N/A 2	-	-
Comorbidities	Hypertension 43 High cholesterol 23 Depression 17 Type II diabetes 12 Sleep apnea 31	-	-
Prior vaccination	Vaccinated 2018-2019 231 Vaccinated 2017-2018 211 Vaccinated 2016-2017 193 Not in prior 3 years 34	-	-

Whole blood samples were acquired from participants prior to and following vaccination (**Figure 1A**). To assess immunogenicity to the influenza HA protein, a hemagglutination inhibition (HAI) assay was performed using two-fold serial dilutions of serum from each participant and turkey red blood cells (RBCs) (**methods**). A numerical HAI value was assigned based on the reciprocal of the final serum dilution at which RBC agglutination inhibition was still observed. For example, if an individual's serum sample inhibited viral-mediated agglutination of RBCs at a dilution of 1:40, but did not inhibit agglutination at a dilution of 1:80, an HAI score of 40 was assigned. The HAI assay was initiated using a 1:5 dilution of serum and tested using a two-fold dilution series of serum. Thus, HAI values vary from 5 to 1280 with an HAI of 5 corresponding to no detectable antibody titer and higher values reflecting higher antibody titers.

To test the specificity of immunogenicity, HAI assays were performed with each of the four vaccine strains prior to vaccination (day 0, or "baseline HAI") and four weeks post-vaccination (day 28, or "response HAI") for all participants (**methods**). However, because the Yamagata strain was not included in the HD formulation the day 28 HAI value was excluded in the serology metrics calculated for those individuals who received the HD vaccine. To account for the different number of strains used in the two formulations we computed average baseline and response serology metrics.

The distribution of baseline HAI values for the four different strains varies between individuals and strains (**Figure S1**) reflecting significant variation in strain-specific pre-existing immunity among individuals. We observed an increase in the distribution of HAI values for all four strains 28 days after vaccine administration (**Figure S2**). High response HAI values for each of the four strains reflect both an immunogenic response to

the administered vaccine and pre-existing immunity. In general, baseline and response HAI values for each strain show positive correlation (**Figure S3**). However, individuals with low baseline HAI values show extensive variation in response HAI.

To quantify the response to vaccination, a seroconversion score was computed for each strain by taking the ratio of the response HAI to the baseline HAI. Seroconversion scores were computed separately for each of the four influenza strains and expressed as $\log_2$ value (i.e. a seroconversion score of 1 corresponds to a twofold increase in HAI value). We find that there is a negative correlation ($\rho \leq -0.2$) between baseline HAI and seroconversion for each of the four strains (**Figure S4**) as high seroconversion scores are generally only observed in those participants with low baseline HAI values. Seroconversion with respect to the Yamagata strain is significantly higher in those individuals who received the SD formulation compared with those receiving the HD formulation, which lacks the Yamagata component, consistent with a specific response to vaccine components (**Figure S5**). Moreover, none of the participants receiving the HD vaccine had a seroconversion score greater than 2 (i.e. a fourfold increase in HAI) for the Yamagata strain supporting the use of a fourfold increase as indicative of a specific antibody response.

To define an average seroconversion score that incorporates information for each strain we computed the average untransformed seroconversion score for the four (SD) or three (HD) strains and expressed the average as a $\log_2$ value (adapted from (Abreu et al. 2020). In general, we treated seroconversion scores as continuous values; however, the HAI assay is known to exhibit significant interassay variation (Stephenson et al. 2007). Therefore, for some analyses we categorized average seroconversion as “low” (< 2, which corresponds to less than a fourfold average change in HAI) or “high” (≥ 2, which corresponds to a fourfold or higher average change in HAI), consistent with the US Food and Drug Administration Guidance (“Guidance for Industry Clinical Data Needed to Support the Licensure of Pandemic Influenza Vaccines” 2007)

Average baseline HAI values, incorporating information from all strains, are negatively correlated with average seroconversion scores ($\rho = -0.34$, p-value = 1.12e-08; **Figure 1B**). Among our study participants we find a significant positive relationship between BMI and seroconversion ($\rho = 0.21$, p-value = 0.0005) (**Figure 1C**) and a slight negative correlation between age and seroconversion ($\rho = -0.13$, p-value = 0.02; **Figure 1D**). Seroconversion does not differ significantly between males and females (F statistic = 3.435, p-value = 0.649).

The majority of study participants received at least one vaccine in the three years prior to the study. However, a small number of participants (n = 34) reported no prior vaccination. We find that these individuals have significantly higher seroconversion scores (t-statistic = -61.8, p-value = 2.2e-16) (**Figure 1E)** consistent with immunologically naive individuals mounting an enhanced response to vaccination.

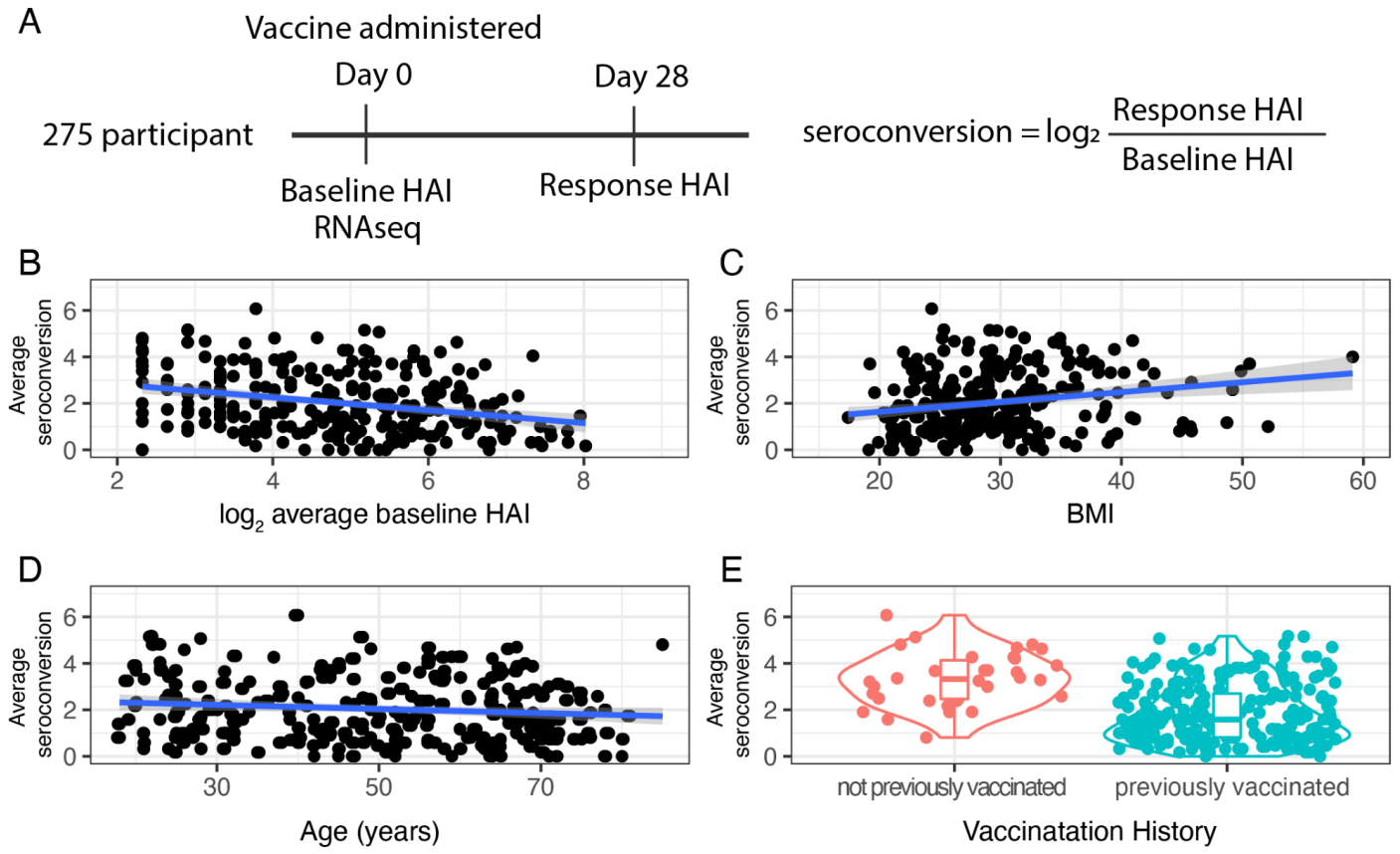


Figure 1. Study design and response to vaccination.

A) 275 participants from the UGA4 vaccination study were selected for whole blood gene expression analysis. HAI for four strains was determined prior to vaccination (baseline HAI) and 28 days (response HAI) after vaccination and used to compute an average seroconversion score. **B**) There is a significant negative relationship between average baseline HAI and average seroconversion (ρ = -0.34, p-value = 1.12e-08). **C**) There is a significant positive relationship between BMI and average seroconversion (ρ = 0.21, p-value = 0.0005). **D**) There is a slight negative relationship between age and average seroconversion (ρ = -0.13, p-value = 0.02). **E**) Those individuals who were not previously vaccinated have higher average seroconversion scores than previously vaccinated individuals.

Differential gene expression analysis

We performed whole transcriptome analysis of whole blood samples for all study participants using RNA sequencing (RNAseq) and processed sequencing data using standard bioinformatic pipelines (**methods**) to generate a table of gene counts (**Table S2**). We obtained a median of 15 million reads per sample (**Figure S6**). Within samples, the distribution of counts per gene is highly skewed as just three genes, encoding hemoglobin, account for almost 50% of the total transcript counts in each sample (**Figure S7**). Therefore, these three genes were excluded from downstream analyses (**methods**). To identify major sources of variation in the data we performed dimensionality reduction analysis. More than 50% of the variance in the data is captured by the first three principal components (**Figure S8**). However, visual inspection failed to identify covariates that were non-randomly distributed across these principal components.

We sought to identify genes that show evidence for differences in expression as a function of study participant physiological (**Table 1**) and serological metrics. Therefore, we performed differential gene expression analysis to identify genes with statistically significant differences in expression (**methods**). Variables that had highly unbalanced categories (i.e. prior vaccination history, and self-reported race) were not considered. We performed differential gene expression analyses using either discrete groups (e.g. for sex) or linear models for continuous variables (e.g. age, BMI, HAI, and seroconversion). We assessed differential gene expression as a function of 1) $\log_2$ transformed average baseline HAI, to identify genes that are differentially expressed with variation in pre-existing immunity, and 2) $\log_2$ transformed average seroconversion, to identify genes whose expression prior to vaccination are predictive of the vaccine response.

When considering the pre-existing immune status of individuals we find only 45 genes that are differentially expressed as a function of baseline HAI (**Figure 2A**). 9 of these genes were uniquely differentially expressed as a function of baseline HAI, while the other 36 were also differentially expressed as a function of different physiological factors. Among the genes that are significantly increased in expression with higher baseline HAI levels are immunoglobulin genes, including *IGLV3-25* (**Figure 2B**), whereas some genes, such as *LDOC1*, show significant negative expression relationships with baseline HAI (**Figure 2C**). Another immunoglobulin gene, *IGLV4-69* (logFC = 4.52, FDR = 7.91e-8), is also increased in expression with baseline HAI. Other immunoglobulin genes are also among the 45 differentially expressed genes, and include *IVLG8-61* (logFC = 3.88, FDR = 1.89e-5), *IGLV4-60* (logFC = 3.32, FDR = 1.59e-3), *IGLV3-10* (logFC = 2.79, FDR = 4.63e-3), and *IGLV3-1* (logFC = 1.71, FDR = 3.1e-2). Among the genes that show the greatest negative relationship to baseline HAI are non-immunoglobulin genes that encode tryptase gamma 1 (*TPSG1,* logFC = -5.82, FDR = 7.91e-8*)* and fibromodulin (*FMOD,* logFC = -5.58, FDR = 1.63e-7*).* Immunoglobulins are expected to be associated with pre-existing immunity and their increased expression may reflect greater plasmablast cell populations in whole blood from individuals with higher baseline HAI. By contrast, the reduced expression of *TPSG1* with increasing baseline HAI may reflect the role of mast cell protease tryptase gamma 1 in mediation of IL-13/IL-4R/STAT6-dependent proinflammatory pathways via T-cell induction (Wong et al. 2002). Quenching of proinflammatory pathways may be required for maintaining a high level of immune response as lower levels of TPSG1 indicate a lower risk of COVID-19 hospitalization (Gaziano et al. 2021). *FMOD* encodes a member of the small interstitial proteoglycans which may play a role in the assembly of the extracellular matrix as well as in regulating TGFβ activity.

When considering differential expression as a function of seroconversion, we identified a set of 84 genes that are uniquely differentially expressed (i.e. show no significant gene expression differences with other participant metrics). A larger set of 741 genes are differentially expressed as a function of seroconversion, but also show significant association with at least one other factor (**Figure 2A**). These genes exhibit both increased (**Figure 2D**) and decreased (**Figure 2E**) expression with increasing seroconversion. The most significantly differentially expressed genes among this gene set include an uncharacterized transcript *ENSG00000273956* (FDR = 7.52e-22), and immunoglobulin genes *IGLV4-69* (FDR = 6.07e-11) and *IGLV8-61* (FDR = 5.43e-7) (**Figure S10)**. Interestingly, *IGLV4-69* has been identified as a component of a monoclonal antibody in a prior vaccination cohort and possesses a broad spectrum binding affinity against a broad range of H3N2 strains, including against the A/Hong Kong/4801/2014 (HK14) vaccine strain used in the 2019-2020 season (Forgacs et al. 2021).

To determine the functions of genes that are differentially expressed depending on baseline HAI and seroconversion score, we performed gene set enrichment analysis (GSEA) of the set of significant genes, (**Figure 2F**). Gene functions that are related to increased baseline HAI include those involved with intracellular transport and immune response. Gene functions that are positively associated with increased seroconversion

include calcium ion import, and regulation of the mitotic cell cycle. Conversely, we find gene functions related to keratin filament and pancreas development to be decreased in expression in individuals with high seroconversion scores. (**Figure 2F)**.

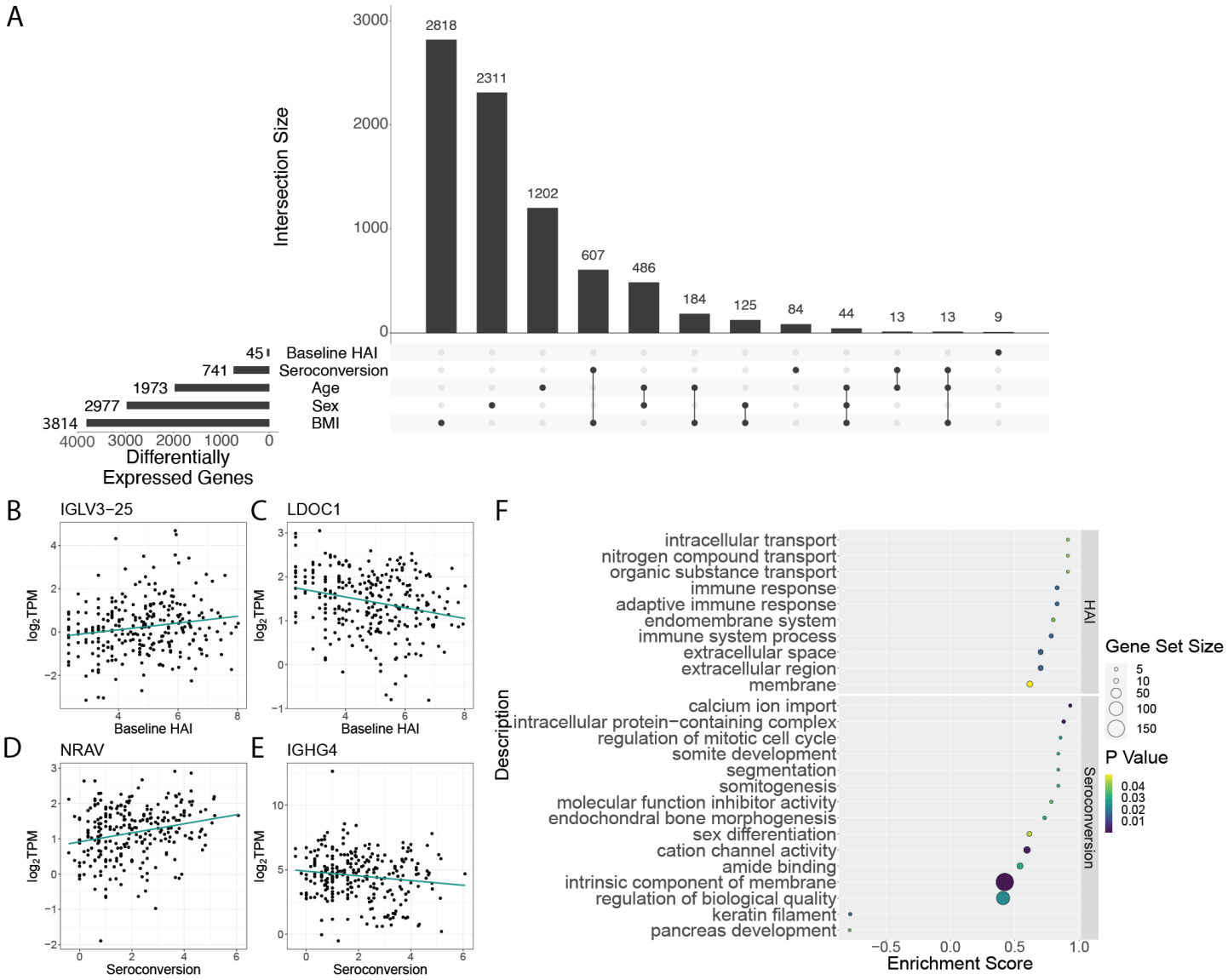


Figure 2. Differential gene expression analysis of pre-existing immunity and vaccine response.

A) Genes that are differentially expressed with variation in baseline HAI and seroconversion. Differential expression analysis was also performed to test the effect of additional participant metrics. Sets of genes that are differentially expressed with two or more variables are connected by lines. Representative genes with **B)** increased and **C)** decreased expression in participants with higher baseline HAI . Representative genes that are **D)** increased and **E)** decreased in participants with higher seroconversion. **F**) Gene function enrichment using GSEA of the 45 genes differentially expressed with baseline HAI (top) and the 741 genes differentially expressed with seroconversion (bottom). Note that the histogram in panel A has been truncated for values less than 9 for visualization purposes.

Interactive effects on differential gene expression

Genes that are differentially expressed as a function of immune status may show differing behaviors depending on other physiological factors. To test for interactive effects between immunological metrics and factors that vary among study participants such as age, BMI, and sex, we used bivariate linear models for both baseline HAI and seroconversion score. We found a number of genes with interactive effects between baseline HAI and sex (1,621), age (911), and BMI (111) (**Figure 3A**). In our bivariate analysis with seroconversion and other covariates, we found the greatest number of interactive effects between seroconversion and age (1,834 genes) (**Figure 3B**). We also found a set of 1,048 genes that show significant interactive effects between seroconversion and BMI, and 840 genes that display significant interaction effects between seroconversion and sex (**Figure 3B**).

We were particularly interested in the interaction between BMI and either baseline HAI or seroconversion as BMI has been identified as one of the major factors associated with increased influenza risk (Neidich et al. 2017). The differential expression of 111 genes with baseline HAI is influenced by BMI (**Figure 3B** and **Figure S11**). For example, *KRT79* increases in expression with higher baseline HAI in overweight individuals, but decreases in individuals with normal BMI (**Figure 3C**). A similar effect of BMI on the relationship between gene expression and baseline HAI is observed for *IGLV4-60* (**Figure 3D**). The differential expression of 1048 genes with seroconversion is influenced by BMI (**Figure 3B** and **Figure S12**). For example, *PCSK1N* increases in expression with seroconversion in overweight individuals, but not individuals with normal BMI (**Figure 3E**). Conversely, *DAAM2* decreases in expression with seroconversion in overweight individuals but not in individuals with normal BMI (**Figure 3F**).

We investigated enrichment of gene sets that show significant interaction between BMI and either baseline HAI or seroconversion. Genes that show interactive effects with BMI and baseline HAI are enriched for protein degradation pathways (**Figure 3G** upper panel). Genes that show interactive effects between BMI and seroconversion are positively enriched for diverse functions including development, metabolism and T cell receptor functions (**Figure 3G** lower panel).

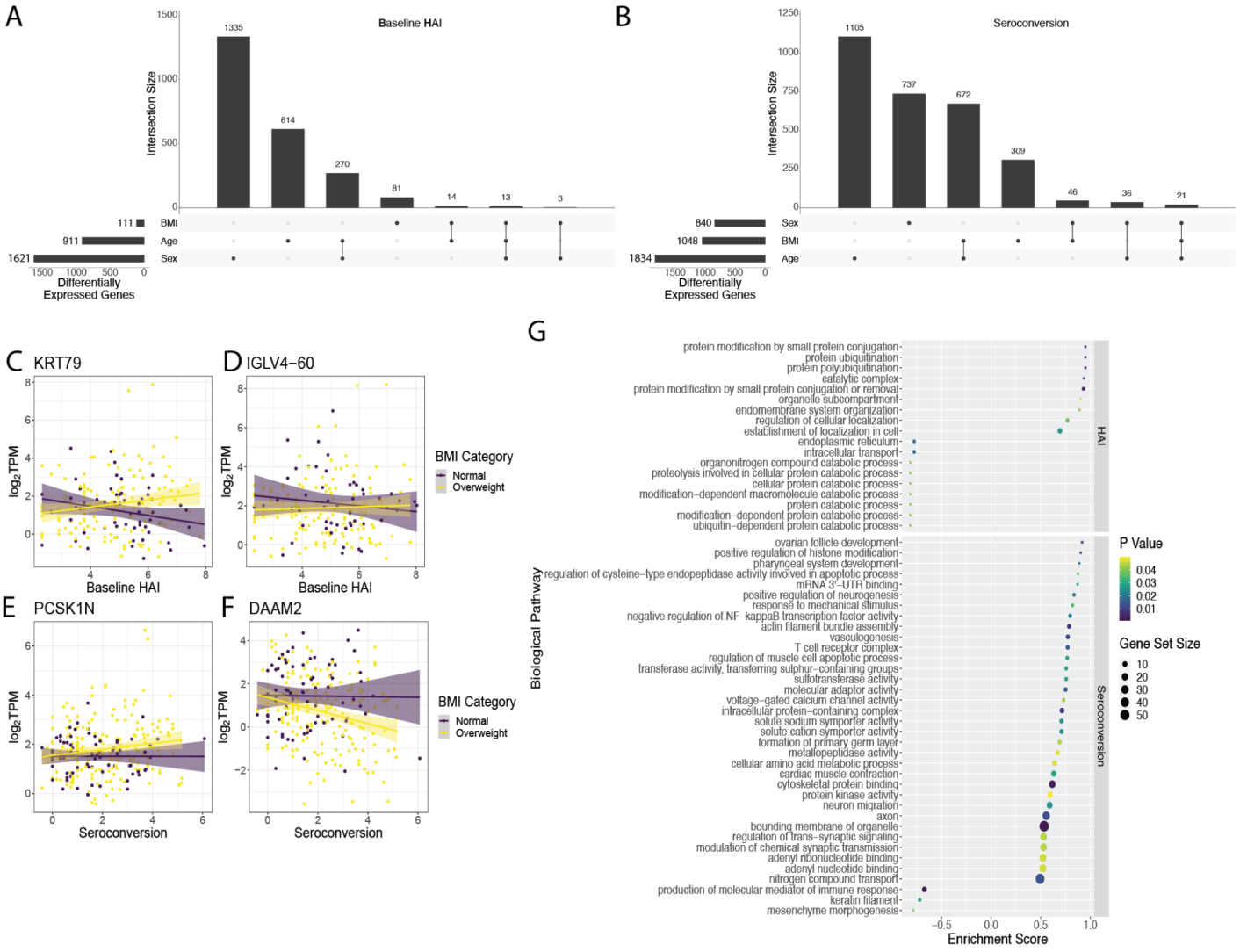


Figure 3. Interactive effects of physiological variables on differential expression with baseline HAI and seroconversion.

Sets of genes showing interactive effects with sex, BMI and age on differential gene expression with **A**) baseline HAI and **B)** seroconversion. Representative genes in which BMI impacts **C)** positive and **D)** negative differential expression with baseline HAI. Representative genes in which BMI impacts **E)** positive and **F)** negative differential expression with seroconversion. Gene function enrichment using GSEA of **G)** 111 genes that show interactive effects with BMI and baseline HAI and **H)** 1,048 genes that show interactive effects with BMI and seroconversion.

Identification of gene expression modules related to serological outcome

We sought to identify molecular processes that show coherent gene expression behavior among individuals. Therefore, we defined sets of genes that show correlated co-expression patterns across individuals using MEGENA (**methods**). The MEGENA method identifies modules of correlated gene expression across samples and defines a hierarchical relationship between modules defined by different levels of stringency. Using the 275 transcriptomes, we identified 489 hierarchically ordered expression modules containing between 10 to 6,324 genes.

We employed different measures based on correlation, differential gene expression enrichment, and a combination of both, to quantify the association between modules and physiological and serological metrics. We use the "Product of Rank" method (**methods**) to compute an aggregated rank from individually ranked modules. The top ten highest ranked modules after enrichment of modules for baseline HAI, response HAI and seroconversion differentially expressed gene (DEG) signatures, are related to the adaptive immune response and immunoglobulin complexes (**Figure 4**) . These modules are strongly enriched for genes that are differentially expressed as a function of response HAI, but show only marginal enrichment for genes that are differentially expressed as a function of baseline HAI and seroconversion (**Figure 4A**). However, we find significant positive correlation between these modules in all three serological scores (**Figure 4B**). The hierarchical organization of 6 of the top ten modules is M17 → M103 → M385 and M17 → M103 → M387 → M706/M707, where M17 is the parent module that includes each subset of smaller modules (**Figure 4C**). Module M387 contains many immunoglobulin genes that show increased expression in subjects with high seroconversion **(Figure 4D)**. The most highly connected, or hub genes, of M387 are the variable chain immunoglobulin genes *IGKV1-5*, *IGKV3-15*, *IGKV3-20*, *IGLV1-40*, and the constant chain immunoglobulin gene *IGKC*. It is noteworthy that a proteomic analysis of influenza haemagglutinin-specific antibodies following H1N1pdm09 influenza A vaccination identified *IGKV3-20* as the dominant light chain variant (Adamson et al. 2017). Modules defined by the M17 parental module are enriched for functions related to the adaptive immune response (**Figure 4E**). Interestingly, these modules are also strongly negatively correlated with age (**Figure 4B**). This points to a potential indirect effect of age on serological outcome manifested in the expression state of these immunoglobulin-enriched modules underlying the observed negative relationship between age and seroconversion (**Figure 1D**).

Additional highly ranked modules are suggestive of other relationships between gene expression and serology. For example, the M31 → M199 module family is enriched for ribosomal functions and is negatively correlated with seroconversion and positively correlated with baseline HAI. The module family M76 → M103, is enriched for cytosolic transport and ferrous iron binding and is positively correlated with baseline HAI.
The most significant module relationships are found with age and BMI (**Figure S13**). The module with the most significant correlation with age is M91 (ρ = -0.41, FDR = 2.37e-10), which contains 73 genes, including the chemokine receptor *CCR7*, T-cell differentiation protein *MAL*, noggin (*NOG*) and leucine rich repeat neuronal 3 (*LRRN3*). M91 also has weak Wnt-signaling (*TLE2, SFRP5, WNT7A, DCHS1*) functionality.
With respect to age association, most modules are enriched for age-specific differential gene expression (**Figure S14A**) and all are strongly correlated with age (**Figure S14B**). Interestingly, the best ranked modules with age are the same as the highest ranked modules based on the serological response (**Figure 4**). The module M17 → M103 → M387 → M706/M707, which is enriched for immunoglobulin functions, is negatively correlated with age and positively correlated with baseline HAI, response HAI, and seroconversion. However, none of these modules are enriched for age-related differential gene expression. On the other hand, the top ranked module M91 (and related module M351) is significantly enriched for differentially expressed genes responding to age. Although M91 is enriched for GO functions related to the elasticity of structural proteins of the extracellular matrix, it involves key regulators related to B- and T-cell responses, such as C-C motif chemokine receptor 7 (*CCR7*), or mal, T cell differentiation protein (*MAL*; **Figure S14D**). Another example of an age-related DEG enriched module is M505. This module is positively correlated with age and does not show any significant correlation with the serological responses. M505 is enriched for (neurite) growth cone development and vesicle transport with α-N-acetylgalactosaminidase (*NAGA*) as a key regulator (**Figure S14E**).

With respect to BMI, the highest ranked module is M76 with ferrous iron binding and cytosolic transport functions. M76 with related modules M303/M304 are highly enriched for genes that are differentially expressed

between previously vaccinated and naive participants (**Figure S15A**). All three modules are weakly positively correlated with seroconversion. Key regulator DDB1 and CUL4 associated factor 12 (*DCAF12*) is an ubiquitin ligase and controls spermatogenesis and T-cell activation (**Figure S15E**) (Lidak et al. 2021). *DCAF12* has also been identified as being differentially expressed between high and low responders after hepatitis B vaccination (Bartholomeus et al. 2018). A second key regulator is anti-apoptosis synuclein alpha (*SNCA*), which is related to Parkinson disease. However, SNCA also plays a role in the response to oxidative stress and has been related to lung function decline (Okeleji et al. 2021); (Singhania et al. 2018). A third key regulator is adiponectin receptor 1 (*ADIPOR1*), which regulates fatty acid catabolism and glucose levels. A second example of a module enriched for BMI DEGs is M47 (**Figure S15D**). This module is positively correlated with both age and BMI but not with serological responses. M47 is enriched for immune effector process functions involving defensins including the key regulator, CEA cell adhesion molecule 8 (*CEACAM8*), a cell-adhesion protein in neutrophils.

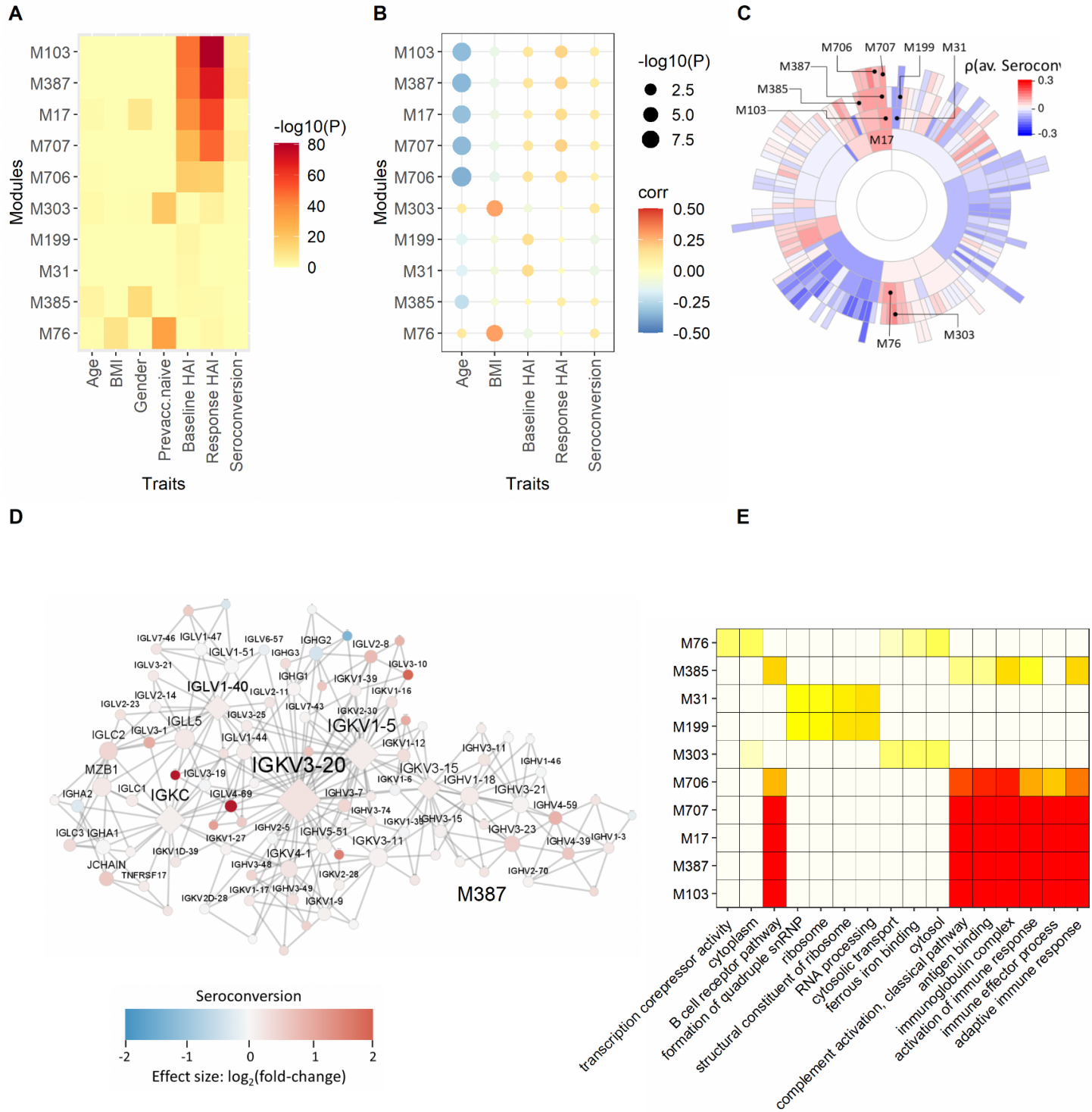


Figure 4. Multi-scale co-expression network analysis.

Using MEGENA we defined gene co-expression network modules associated with participant metrics and immune status. The 10 best ranked modules are shown using a combined rank based on the correlations

related to serological effects and outcome (average baseline, average seroconversion and average response). **(A)** The module enrichment for DEGs are depicted in this heatmap. **(B)** The dot-plot shows the correlation coefficients and corresponding P-value of module / trait correlations. **(C)** A sunburst plot with the hierarchical structure of the module. **(D)** Module M387 enriched for immunoglobulin functionality . Node colors indicate fold change with respect to high versus low average seroconversion, red for increased expression in subjects with high seroconversion compared to subjects with low seroconversion, blue refers to decreased gene expression between these two groups of subjects. **(E)** The functional enrichment of the corresponding modules for GO functions are shown.The hierarchical relationship of the modules shown in **(A)** and **(B)** are as follows: M17 → M103 → M385; M17 → M103 → M387 → M706/M707; M31 → M199; M76 → M303.

Prediction of vaccine response

We tested the predictive value of physiological information and gene expression with respect to seroconversion. For this purpose, we generated random forest models using different combinations of three distinct classes of features: 1) physiological metrics, 2) gene expression of the 741 genes that are differentially expressed as a function of seroconversion, and 3) co-expression modules. We initially attempted to predict quantitative seroconversion scores, but determined that all models performed poorly. Therefore, we classified individual seroconversion scores as high (average seroconversion ≥ 2, n = 124) or low (average seroconversion < 2, n = 151) for the purpose of prediction.

When used to predict seroconversion, we find that a random forest model generated using only physiological metrics results in the lowest out-of-bag (OOB) error rate. The random forest model generated using only differentially expressed genes yields the highest OOB error rate whereas the combined data results in an intermediate error (**Figure 5A**). We then evaluated each model's average predictions by using a receiver operating characteristic (ROC) curve and calculating the resulting area under the curve (AUC). The random forest model generated using only physiological data yielded the highest average AUC of 0.69, the model generated using only expression data yielded the lowest average AUC of 0.61, while the model generated using both yielded an AUC of 0.68 (**Figure 5B**).

To identify the specific instances for which our random forest models fail, we generated confusion matrices using the average predictions of each of the three datasets (**Figure 5C**). All three random forest models have difficulty correctly categorizing high seroconversion, often mistakenly categorizing them as low. However, when assessed using the test dataset, the random forest model generated using only physiological data accurately classifies 53.3 of the 93 high seroconversion individuals on average compared to the models generated using only expression data (41.4 of 93 correctly assigned) and clinical data paired with expression data (44.3 of 93 correctly assigned) respectively.

We performed cross validation analyses for each of the three random forest models to assess the reproducibility of the model on unseen data as well as to identify the optimal number of features for each model. We find that for models that contain physiological data the error is lowest when less than 5 features are used (**Figure 5D**). By comparison, the random forest generated using only gene expression data requires almost 1,000 independent features to minimize predictive error. This suggests that physiological information is more informative about serological response than gene expression. For each random forest model, we calculated and ranked the average importance value for each feature (**Figure 6**). When using a model with only physiological metrics to predict seroconversion prior vaccination status is the most important variable (**Figure 6A**), but BMI, baseline HAI, sex and age also contribute to model performance and reproducibility as measure using the GINI coefficient (**Figure 6B**). When using only differentially expressed genes we find that removal of individual genes has incremental effects on model performance as measured by the decrease in

model accuracy (**Figure 6C**) and GINI coefficient (**Figure 6D**) as do models that combine physiological metrics and differentially expressions genes (**Figure 6E** and **Figure 6F**). This is consistent with no single gene having an expression value that is especially information of seroconversion.

We further investigated whether prediction seroconversion outcome would improve with the inclusion of the aggregated module data generated using MEGENA. Combination of one of principal components 1-5 of the module data with the physiological data and DEGs yields a higher OOB error rate and decreased AUC compared to random forest models generated with either physiological data only or physiological data with DEGs (**Figure S16A**). A cross-validation analysis shows again that each model yields the lowest cross-validation error rate when run using <5 features (**Figure S16B**). Similarly, ranking variables by importance in this set of random forest models consistently indicates previous vaccination status to be the most important variable in predicting seroconversion category (**Figure S17**). Of the clinical variables, baseline HAI, BMI, and previous vaccination status consistently are in the top 10 variables for this set of random forest models. While no modules appear in the top 10 most important features across multiple random forest models in this set, we do see the genes *C8B*, *IGHG4*, *KIAA0319L*, and *LGSN* show up in the top 10 features in multiple random forest models in this set.

When combining physiological features with gene expression, previous vaccination status, BMI, and baseline HAI remain highly predictive along with genes and lncRNAs, such as *CACNB4*, *CD177*, *GNAL*, *GZMH*, *IGHG4*, *KIAA0319L*, *MYOM2*, *PTP4A3*, *TUBB8B, and ENSG00000259352*. Half of the ten best ranked molecular features have immune system relevant functions. *C8B* and *LGSN* are conserved as important factors in the random forest model that combine physiological features, gene expression and module PCs but are not highly ranked in the models without co-expression modules. Overall, the main biological functions of these genes are related to immune processes.

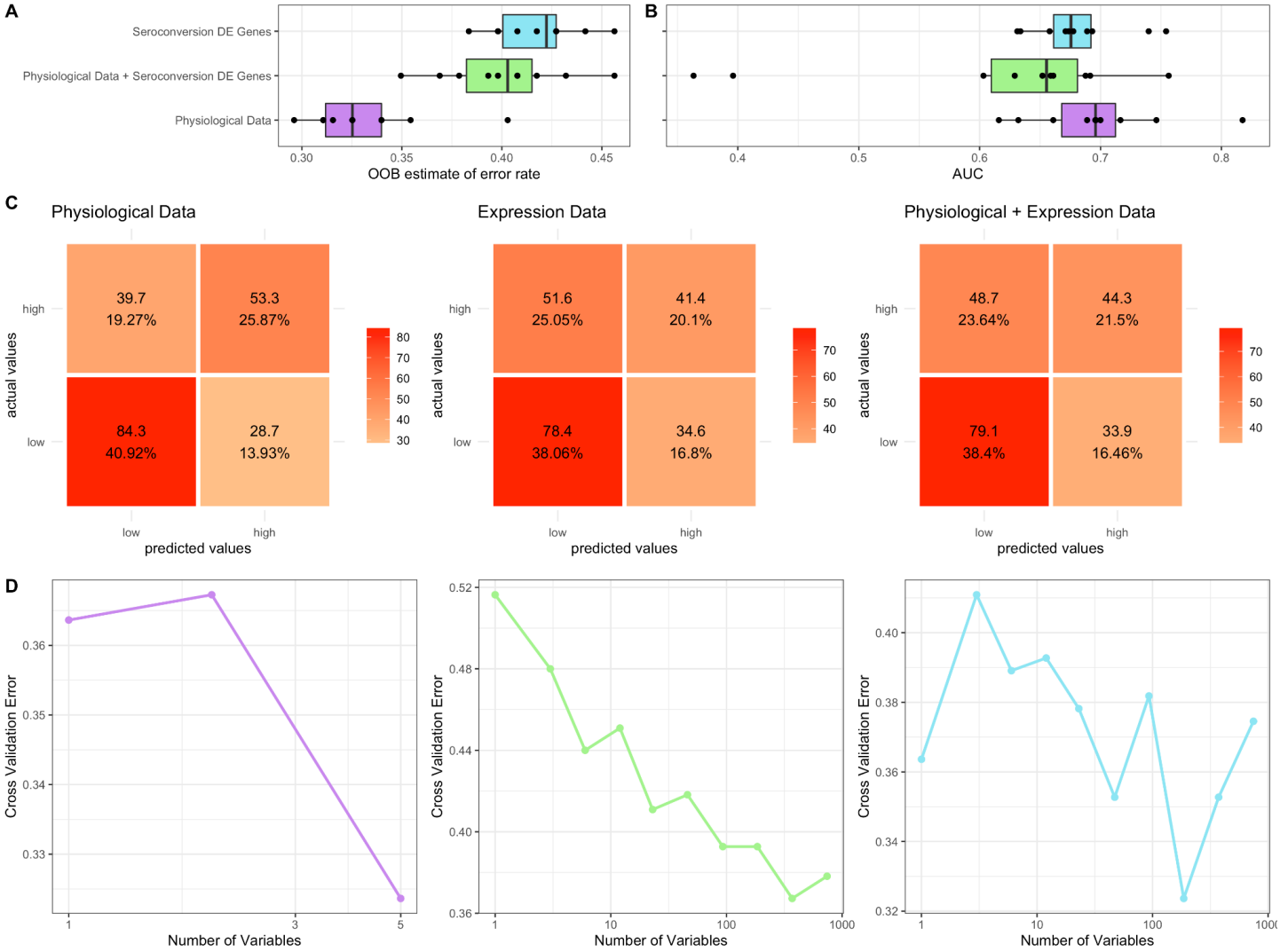


Figure 5. Prediction performance of random forest models generated using physiological and differential gene expression data to predict seroconversion.

The average **(A)** out-of-bag (OOB) error rate and **(B)** average area under the curve (AUC) values generated using 10 iterations of random forest models constructed using physiological data, expression data, and both combined. **(C)** The confusion matrices displaying true and false positives and negatives for the random forest models using physiological data (left), expression data (middle), and both combined (right). **(D)** A cross validation analysis plotting a varying number of variables for the physiological data (left), expression data (middle), and both combined (right) random forest models against the resulting cross validation error.

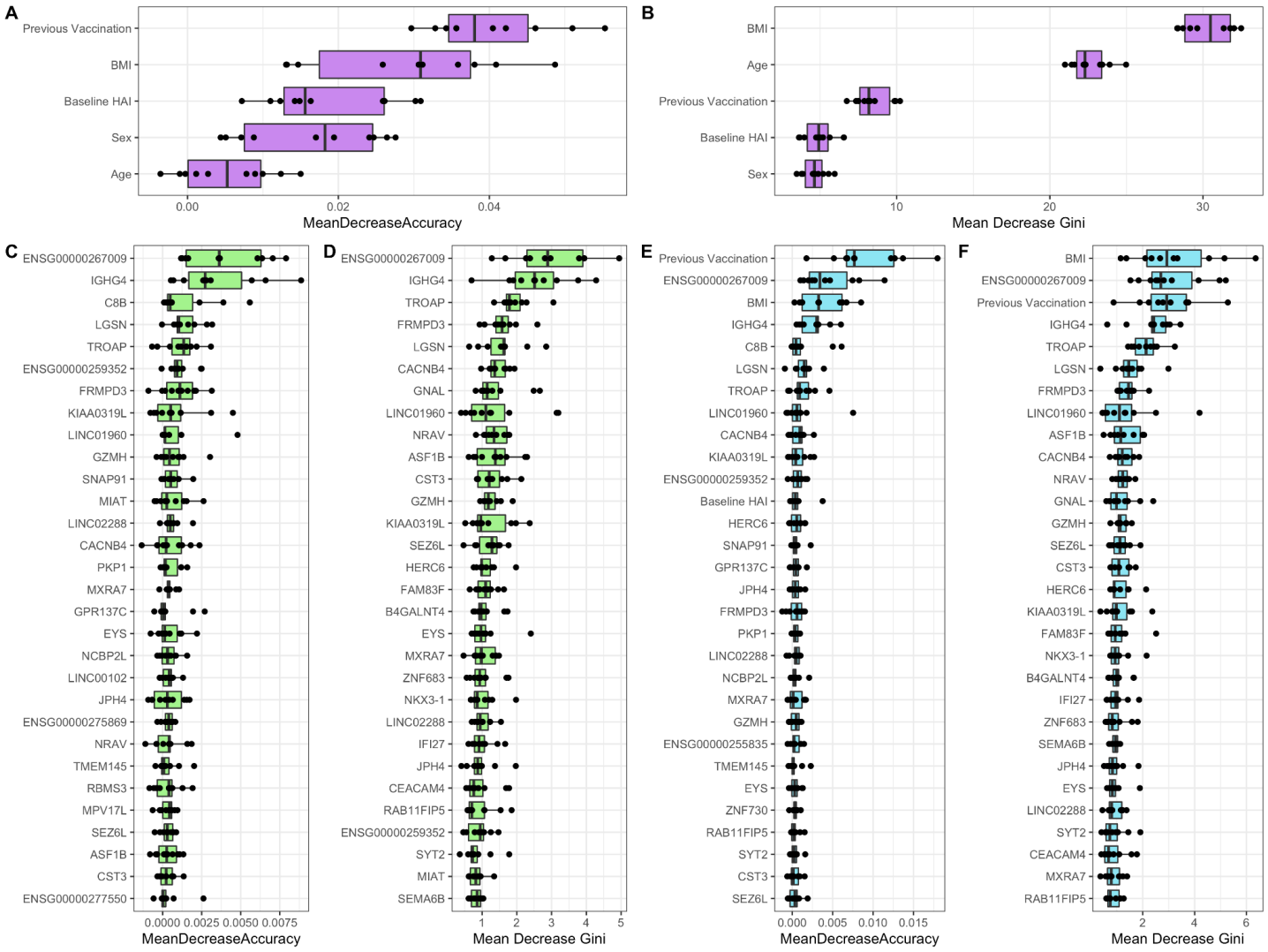


Figure 6. Identification of factors underlying random forest model performance.

Variable importance measured through mean decrease accuracy and mean decrease in Gini coefficient for random forest models generated using participant **A)** and **B)** physiological metrics, **C)** and **D)** the 741 DEGs as a function of seroconversion, and **E)** and **F)** physiological metrics and DEGs combined.

Discussion

The immunological response to the seasonal influenza vaccine exhibits significant heterogeneity among individuals. In this study, we addressed two key questions with respect to the observed variation: (1) to what extent do whole blood gene expression profiles and physiological variables relate to serological status prior to, and following, vaccination, and (ii) are gene expression and physiological metrics predictive of vaccine response.

We first assessed the effect of physiological factors on serological status. As expected, the response to the vaccine is strongly impacted by existing immunity. We quantified existing immunity to each of the four influenza strains included in the vaccine and observed significant variation among individuals. Among those individuals with higher existing immunity we detect a muted increase in immune status following vaccination.

Individuals who had not been vaccinated in the prior three years show a much greater response than those who received at least one influenza vaccine in the prior three years. Surprisingly, we detect a trend of higher BMI associated with increased vaccine response. This could reflect either systematically decreased pre-existing immunity, or a greater degree of immunological priming, in higher BMI individuals.

Gene expression profiling revealed a large number of genes that are differentially expressed genes with physiological factors including BMI (3,814 genes), sex (2977 genes), and age (2,977 genes). By contrast, only 45 genes are differentially expressed as a function of baseline HAI. Surprisingly, a much larger number of genes are differentially expressed with seroconversion (741) although only 11% (84) of these are not associated with one of the physiological factors. This discrepancy suggests that whole blood gene expression profiles obtained prior to vaccination are more informative of the response to vaccination than the pre-existing immunological state of an individual.

Among genes that are increased in expression with increasing seroconversion are several immunoglobulin genes. As gene expression is assayed prior to vaccination this may indicate that those individuals who will mount the greatest serological response to the vaccine are predisposed to do so as a result of increased pre-existing immunological activity (Aw, Silva, and Palmer 2007); (Frasca et al. 2020). It is possible that these expression values may be influenced by imprecise mapping of RNAseq reads to variable gene segments of immunoglobulins. However, we believe that this is unlikely as visual inspection of aligned reads for a representative example (*IGLV4-69*) indicates high quality mapping and BLAST searches of *IGLV4-69* mapped reads against the reference genome identify this feature as the most likely source. Our findings suggest the interesting possibility that enhanced vaccine response may result from either increased plasmablast subpopulations or enhanced expression of specific immunoglobulins. Testing these hypotheses requires additional analyses, possible through variable sequence reconstruction using RNAseq reads (Mandric et al. 2020; Blachly et al. 2015), and cell profiling of participants.

Using a network-based approach we identified co-expression modules that were significantly associated with serological status. As with our gene-level analysis, we identified modules involving immunoglobulins. Module M387 with hub genes *IGKV3-20*, *IGKV3-20*, *IGLV1-40*, *IGKC*, and *IGKV3-15*, together with modules M17, M103, M706 and M707 are positively correlated with baseline HAI, response HAI and seroconversion. These modules are functionally enriched for antigen binding, immunoglobulin complex, and activation of adaptive immune response. We further observed negative correlations between these immunoglobulin modules and age, which may reflect reduced immunity with age.

Central to our analysis was the identification and assessment of factors for the prediction of seasonal influenza vaccine response. For this purpose we employed a random forest approach for both feature selection and modeling. Consistently, the most important physiological predictive features are baseline HAI, BMI, and prior vaccination status. When combining physiological features with gene expression, these three features remain highly predictive along with genes and lncRNAs, such as *CACNB4*, *CD177*, *GNAL*, *GZMH*, *IGHG4*, *KIAA0319L*, *MYOM2*, *PTP4A3*, *TUBB8B, and ENSG00000259352*. About half of the ten genes with high predictive value have immune system functions. *CACNB4* is the calcium voltage-gated channel auxiliary β_4-subunit, responsible for controlling Ca^{2+} influx. *CACNB4* regulates the transcriptional activity of the T-cell factor promotor (TCF), further downregulates the Wnt/β-catenin signaling pathway and inhibits cell proliferation(Rima et al. 2017). Thus, *CACNB4* may play a role in the modulation of immune system responses. *CD177* is a neutrophil-specific cell surface glycoprotein that plays a role in neutrophil activation(Marino et al. 2022). Granzyme H (*GZMH*) is a serine protease. It plays a role in the cytotoxic type of innate immune response by inducing cell death. *IGHG4* codes for the immunoglobulin heavy constant gamma 4 chain and is involved in many processes of the immune system. Protein tyrosine phosphatases, such as *PTP4A3,* are cell

signaling molecules that play regulatory roles in a variety of cellular processes. *PTP4A3* is further a prognostic marker for tumor growths and metastasis. Tubulins, such as tubulin beta 8B (*TUBB8B*) are responsible for microtubule cytoskeleton organization and mitotic cell cycle. They are also key components for the formation of proplatelets from megakaryocytes and during platelet activation(Cuenca-Zamora et al. 2019). Despite the predictive value of gene expression we find that response prediction using exclusively physiological information performs better than models that include gene-expression, either exclusively or additionally.

Overall, our study further highlights the importance of physiological variables on the vaccine response, and identified key genes and co-regulatory networks associated with the individual vaccine response. Whereas physiological factors possess the best power to predict vaccination outcome, in the absence of verified physiological information, the expression of key genes may serve as an informative substitute when this information is not available.

Methods

UGA4 study

Study design

Participants were enrolled at the University of Georgia Clinical and Translational Research Unit (Athens, Georgia) (IRB #3773) from September 2019 to February 2020. All volunteers were enrolled with written, informed consent and excluded if they already received the seasonal influenza vaccine. Other exclusion criteria included acute or chronic conditions that would put the participant at risk for an adverse reaction to the blood draw or the flu vaccine (e.g., Guillain-Barré syndrome or allergies to egg products), or conditions that could skew the analysis (e.g., recent flu symptoms or steroid injections/medications). All participants received a commercially available seasonal influenza vaccine, Fluzone™ (Sanofi Pasteur), which is a split-inactivated vaccine derived from influenza viruses propagated in embryonated chicken eggs. Most participants received a standard dose, quadrivalent vaccine which was formulated with 15µg HA per strain of A/H1N1 (A/Brisbane/02/2018), A/H3N2 (A/Kansas/14/2017), B/Victoria (B/Colorado/6/2017-like strain), and B/Yamagata (B/Phuket/3073/2013). Some participants 65 years and older chose the high-dose vaccine which was a trivalent composition lacking a B/Yamagata strain, but formulated with 60µg HA/strain of the others.

HAI assay

Hemagglutinin inhibition (HAI) assays were performed with serum from each participant pre-vaccination (day 0) and post-vaccination (day 28). Sera was used at a starting concentration of 1:10 following treatment with a receptor-destroying enzyme (RDE) (Denka Seiken) to inactivate non-specific inhibitors. RDE-treated sera (25 µL) were serially diluted in PBS two-fold across 96-well V-bottom microtiter plates to column 11, leaving the last column without sera as a negative control. Similarly treated positive control ferret sera was included on some plates. An equal volume of influenza virus (25 µL), adjusted beforehand via hemagglutination (HA) assay to a concentration of 8 hemagglutination units (HAU) per 50 µL, was added to all wells, plates mixed by agitation, and then incubated at room temperature for 20 minutes. Next, 0.8% turkey red blood cells (Lampire Biologicals) in PBS were added, plates mixed by agitation, and then incubated at room temperature for 30 minutes. The HAI titer was determined by the reciprocal dilution of the last well that contained non-agglutinated RBCs, and a value of 5 was assessed in cases where no HAI was detectable.

RNA sequencing

Library preparation and sequencing

We obtained 275 whole blood samples from the UGA4 study previously described. Each tube contained 2.5mL of blood that was collected prior to vaccination (day 0) into PAXgene Blood RNA Tubes according to the manufacturer's protocol (PreAnalytiX, Hombrechtikon, Switzerland). Before freezing, blood was incubated in the tube's proprietary reagent at room temperature for a minimum of two hours to ensure stabilization of intracellular RNA. Total RNA was isolated from each blood sample in randomized batches using the PAXgene Blood miRNA Kit (QIAGEN, Hilden, Germany) according to the manufacturer's recommendations and stored at -80°C in a LoBind twin.tec 96-well PCR plate (Eppendorf, Hamburg, Germany). All samples were quantified with the Qubit BR RNA Assay on a Qubit 2.0 Fluorometer (ThermoFisher Scientific, Waltham, MA) and their fragment size distribution was measured with an RNA Screentape on a 4200 TapeStation System (Agilent, Santa Clara, CA). Sample yield ranged considerably (min: 0.6ug, max: 24ug) but all samples had sufficient input mass for downstream library prep and RIN values were consistently >7.0.

Sequencing libraries were prepared using the NEBNext Poly(A) mRNA Magnetic Isolation Module and NEBNext Ultra II RNA Library Prep Kit for Illumina (NEB, Ipswich, MA) according to the manufacturer's protocols using either 1ug of input RNA or, in the case of several low-yield samples, the maximum possible input mass. AMPure XP Beads (Beckman Coulter, Brea, CA) were used in place of the supplied NEBNext Sample Purification Beads. Libraries were prepared in 3 batches with two negative library prep controls in each batch to monitor for reagent or sample-to-sample contamination. Each library was barcoded with NEBNext Multiplex Oligos for Illumina (NEB, Ipswich, MA) in a rotating scheme, which ensured that each pool had a unique set of barcodes. Each library was quantified with the Qubit dsDNA HS Assay on a Qubit 2.0 Fluorometer and fragment size was measured with an HSD1000 Screentape (Agilent, Santa Clara, CA) on a 4200 TapeStation System. All libraries made from experimental samples were of sufficient molarity for pooling and sequencing. Libraries were pooled with equimolar input at either 2nM or 3nM into sets of 96 libraries for sequencing on a NovaSeq 6000 with the S1 2x150 - 300 Cycle configuration (Illumina, San Diego, CA) at the Genomics Core Facility (Center for Genomics and Systems Biology, New York University).

Upstream processing

Sequence reads were aligned to the human transcriptome using the nf-core/rnaseq nextflow pipeline (Patel et al. 2022) with default parameters. This pipeline aligned reads to the human genome (GRCh37) using STAR (Dobin et al. 2013) followed by BAM-level quantification with Salmon (Srivastava et al. 2020) to generate a feature counts table. All gene expression information and fastq files have been submitted to GEO.

After quantifying human transcripts at the gene level, the count data was processed to keep only one sample per individual with the highest number of gene counts. Additionally, counts from pseudogenes, as annotated by biomaRt in the hsapiens_gene_dataset Ensembl database, were filtered out from any downstream analyses along with the three genes ENSG00000188536, ENSG00000244734, and ENSG00000206172, which correspond to highly expressed hemoglobin genes. TPM values were calculated for all remaining genes using the average gene length of the individual transcripts encoded by their corresponding gene.

Univariate differential expression analyses

Differentially expressed genes as a function of seroconversion score, baseline HAI, age, gender , BMI , race, month vaccinated, and previous vaccination status were separately identified using edgeR v. 3.30.3 (Robinson, McCarthy, and Smyth 2010) with a FDR cutoff of 0.05. For seroconversion score, age, and BMI a fit spline with

1 degree of freedom was used to construct the design matrix while for all other variables, discrete groups were used. Only genes with at least 10 counts-per-million (CPM) in 70% of samples in the smallest group were kept for differential expression analyses. The remaining counts in each sample were normalized by trimmed mean of M-values (TMM) (Robinson and Oshlack 2010) and dispersions were estimated and fit to a quasi-likelihood negative binomial generalized log-linear model (glmQLFit) with robust set to true. UpSet plots (Lex et al. 2014) used to display intersections of differentially expressed genes were plotted using the R package UpSetR v1.4.0.

GSEA analyses for the sets of genes differentially expressed as a function of seroconversion score and baseline HAI were done using gene lists ranked by a decreasing FC value. GSEA were conducted using the two sets of ranked gene lists individually using the human annotation org.Hs.eg.db v3.11.4, and the R package clusterProfiler v.3.16.1 (Yu et al. 2012) with the function gseGO with minGSSize = 3, maxGSSize = 800, and a FDR cutoff of 0.05. Redundant GO terms were removed using both the simplify function from clusterProfiler with a similarity cutoff of 0.7 (baseline HAI) and 0.5 (seroconversion) and manual curation.

Bivariate differential expression analyses

For seroconversion score and baseline HAI, bivariate differential expression analyses were done using edgeR v. 3.30.3 (Robinson, McCarthy, and Smyth 2010) with a FDR cutoff of 0.05 by constructing a design matrix with one of the two immune terms paired with age, gender, and BMI. A full rank design matrix was unable to be constructed when pairing initial HAI paired with sex, so those pairings were excluded from downstream analyses. Normalization, dispersion estimation, and linear model fitting were done the same way as for the univariate differential expression analyses. UpSet plots (Lex et al. 2014) used to display intersections of differentially expressed genes were plotted using the R package UpSetR v1.4.0.

GSEA analyses for both DEGs of seroconversion score and initial HAI as a function of BMI were done using a list of genes ranked by a decreasing FC value. GSEA was conducted using both sets of ranked gene lists individually using the human annotation org.Hs.eg.db v3.11.4, and the R package clusterProfiler v.3.16.1 (Yu et al. 2012) with the function gseGO with minGSSize = 3, maxGSSize = 800, and FDR cutoff of 0.05. Redundant GO terms were simplified using both the simplify function from clusterProfiler with a similarity cutoff of 1 (baseline HAI) and 0.4 (seroconversion) and manual curation.

Multi-scale network analysis

We performed Multiscale Embedded Gene Co-Expression Network Analysis (MEGENA) [1] to identify host modules of highly co-expressed genes in influenza infection. The MEGENA workflow comprises four major steps: 1) Fast Planar Filtered Network construction (FPFNC), 2) Multiscale Clustering Analysis (MCA), 3) Multiscale Hub Analysis (MHA), 4) and Cluster-Trait Association Analysis (CTA). The total relevance of each module to influenza virus infection was calculated by using the Product of Rank method with the combined enrichment of the differentially expressed gene (DEG) signatures as implemented: $G_j = \prod_j g_{ji}$, where, g_{ji} is the relevance of a consensus $j$to a signature i; and g_{ji} is defined as $\left(\max_j (r_{ji}) + 1 - r_{ji} \right) / \sum_j r_{ji}$, where r_{ji} is the ranking order of the significance level of the overlap between module j and the signature.
The correlation between modules and traits was performed using Spearman's correlation.

Identification of enriched pathways and key regulators in transcriptome modules

The biological functions of identified modules in this study were assessed by enrichment analysis for established pathways and pathway (gene) signatures), including the gene ontology (GO) (Gene Ontology Consortium 2015) biological processes (BP) category and MSigDB (Subramanian et al. 2005) canonical pathways (C2.CP) (GSEA, 2022). Enrichment analysis was performed using Fisher's Exact Test (FET; inhouse and the hypergeometric test from the Category R-package).
The analysis to identify key regulators takes as input a set of genes (*G*) and a co-expression network. The objective is to identify the key regulators for the gene sets with respect to the given network. This approach first generates a subnetwork *NG*, defined as the set of nodes in *N* that are no more than h layers away from the nodes in *G*, and then searches the *h*-layer neighborhood (*h = 1*, …, *H*) for each gene in *NG* ($HLN_{g,h}$) for the optimal h^*, such that

$$ES_h^* = \max(ES_{h,g}) \; \forall g \in N_g, \; h \in \{1, \ldots, H\}$$

where $ES_{h,g}$ is the computed enrichment statistic for $HLN_{g,h}$. A node becomes a candidate driver if its HLN is significantly enriched for the nodes in *G*. Candidate drivers without any parent node (i.e., root nodes in directed networks) are designated as global drivers and the rest are local drivers. To identify the system specific key regulators in their coexpression networks, the corresponding SRGs have been used as gene set *G*.

Random forest models

Random forest models to predict seroconversion category (low: seroconversion score <2; high: seroconversion score >=2) were done using the R package randomForest v4.6-14 using a combination of (1) clinical variables including initial HAI, BMI, age, race, sex prevaccination status, and month vaccinated, (2) the 741 genes identified to be differentially expressed as a function of seroconversion, and (3) one of the first five principal components of the identified expression modules.

A total of 10 iterations for each input data set were used to generate random forest models. Samples were first split such that 75% of samples were used in the training set with the remaining 25% being used as a test set. The tuneRF function, with the options ntreeTry=500, stepFactor=1.5, improve=0.01, and using a start mtry value of a third of the number of total samples.were used to identify the optimal mtry value for the random forest model. From this, a random forest model was generated using the optimal mtry value with 500 trees grown.

The average out of bag (OOB) error rate and area under the curve (AUC) for the plotted ROC curves of the 10 random forest iterations for a given dataset were used to assess the efficacy of each model in predicting seroconversion category. Cross validation using the test dataset was done using a mtryStart value equal to a third of the number of samples and a cv.fold value of 100. The average importance values for each set of features for each random forest model was identified by calculating mean decrease accuracy and the mean decrease in Gini index.

Data availability

All computer code to reproduce all figures and analysis in this study is available on Github (https://github.com/GreshamLab/CIVR_HRP_Day0). RNAseq data has been submitted to GEO.

Acknowledgements

We thank members of the Gresham and Ghedin labs. Funding provided by NIAID (75N93019C00052), R21AI149013 (CVF), R01140766 (DG) and NIGMS (R01GM107466 and R01GM134066) (DG). This work was funded in part by the Division of Intramural Research (DIR) at NIAID/NIH (EG, MC). This work was supported in part through the computational resources provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai (https://labs.icahn.mssm.edu/minervalab). This work utilized the computational resources of the NIH High Performance Computing (HPC) Biowulf cluster (http://hpc.nih.gov).

Supplementary Figures

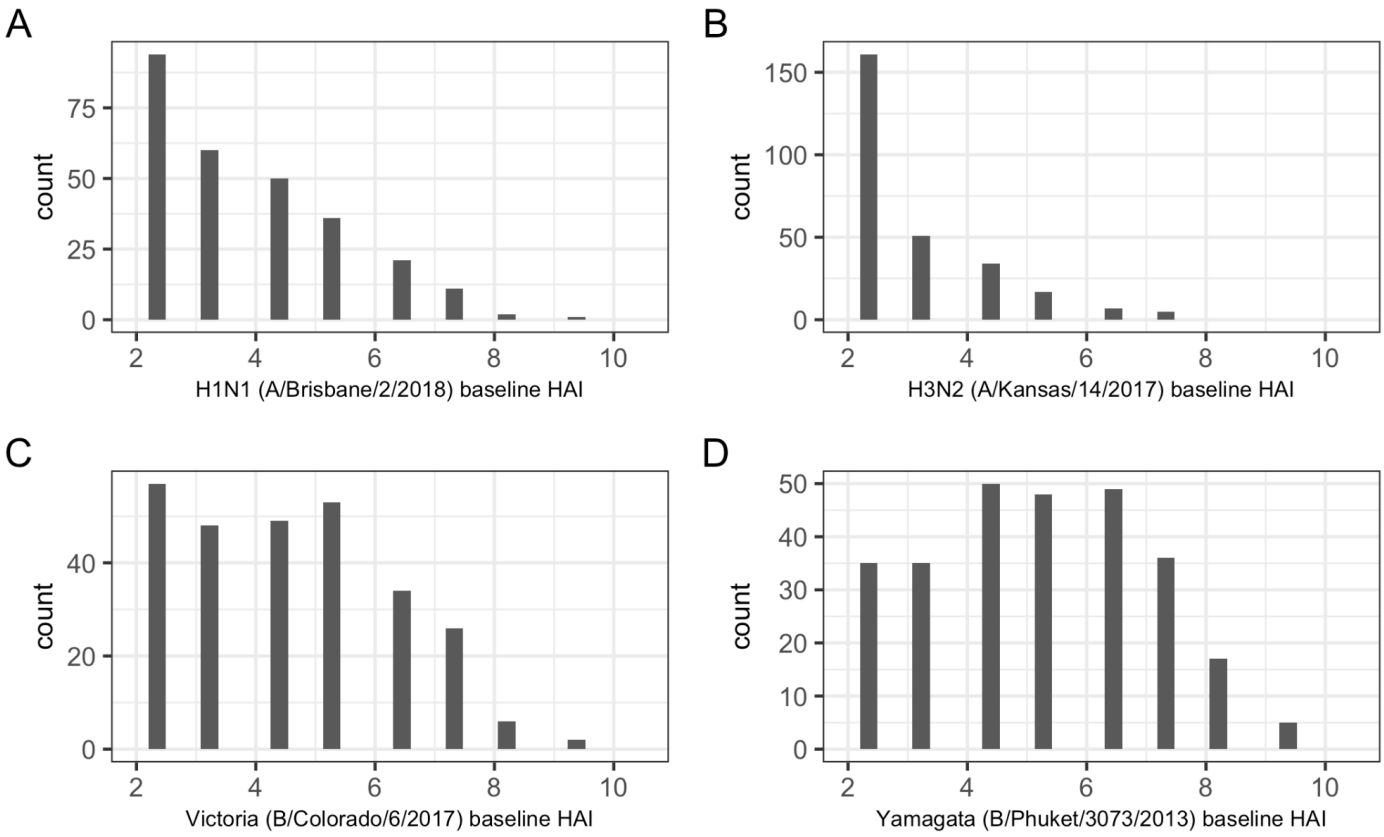


Figure S1. **Distribution of baseline HAI for each influenza strain among study participants.** The number of individuals (count) with the corresponding HAI value prior to vaccination for the **A**) H1N1, **B**), H3N2, **C**) Victoria, and **D**) Yamagata strains included in the vaccine formulations. Values for all 275 individuals are reported for each of the four strains.

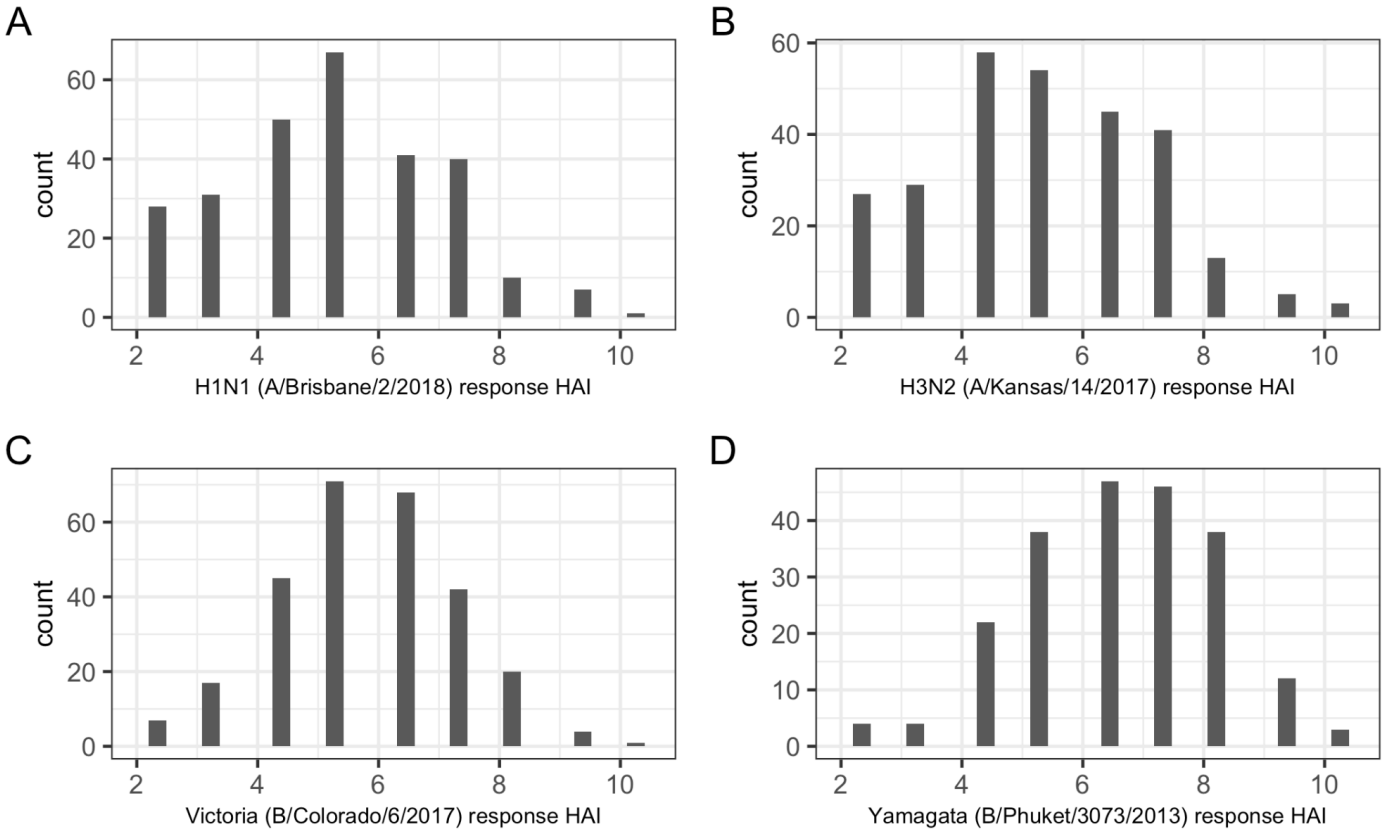


Figure S2. **Distribution of response HAI for each influenza strain among study participants.** The number of individuals (count) with the corresponding HAI value 28 days after vaccination for the **A**) H1N1, **B**), H3N2, **C**) Victoria, and **D**) Yamagata strains included in the vaccine formulations. Values for all 275 individuals are reported for each of the three of the strains, but only the 214 individuals who received the standard dose formulation containing the Yamagata component are shown in panel D.

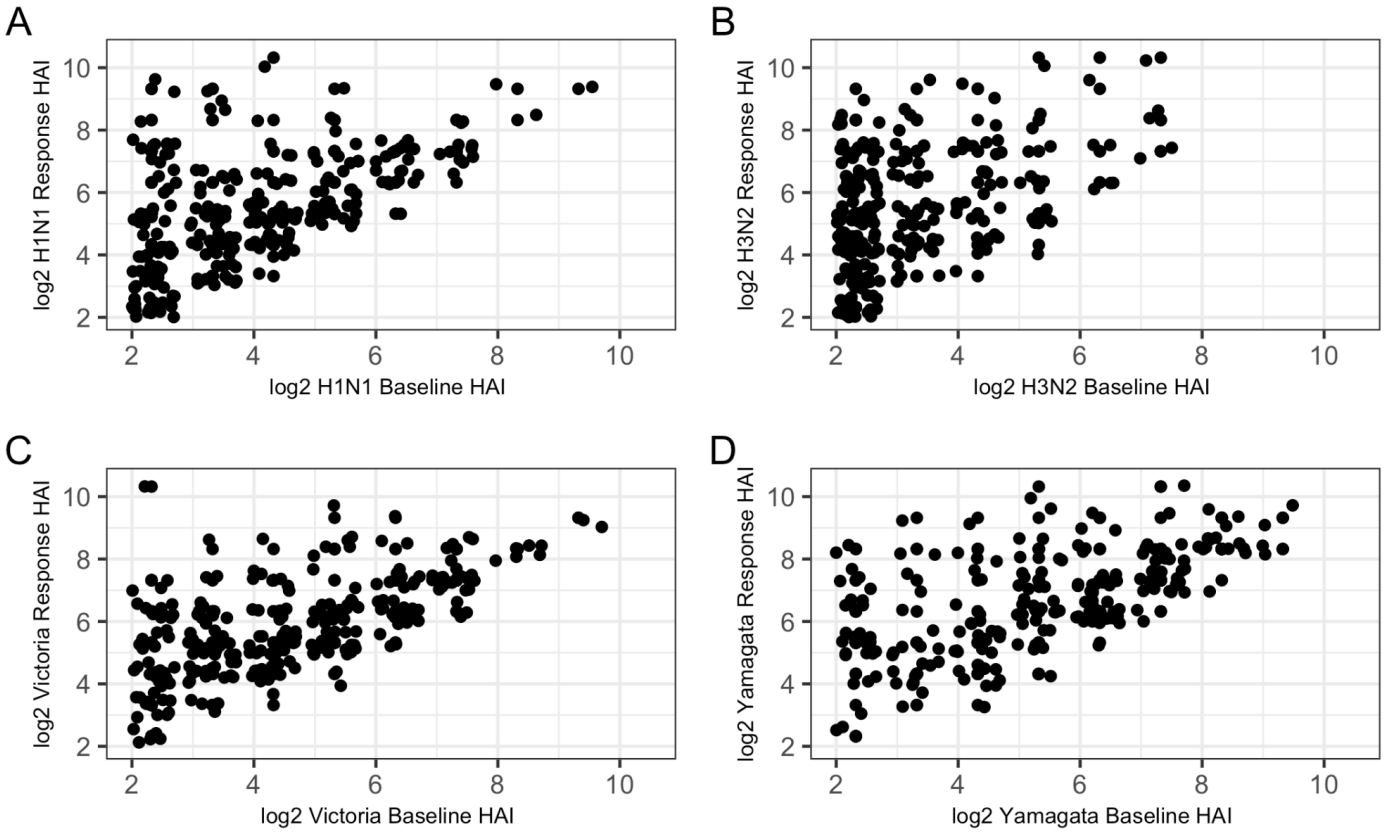


Figure S3. **Relationship between baseline HAI and response HAI for each strain.** The baseline HAI, measured prior to vaccination and the corresponding HAI value 28 days after vaccination for the **A**) H1N1, **B**), H3N2, **C**) Victoria, and **D**) Yamagata strains included in the vaccine formulations. Values for all 275 individuals are reported for each of the three of the strains, but only the 214 individuals who received the standard dose formulation containing the Yamagata component are shown in panel D. A small amount of random noise was added to each value for visualization purposes to avoid overplotting points.

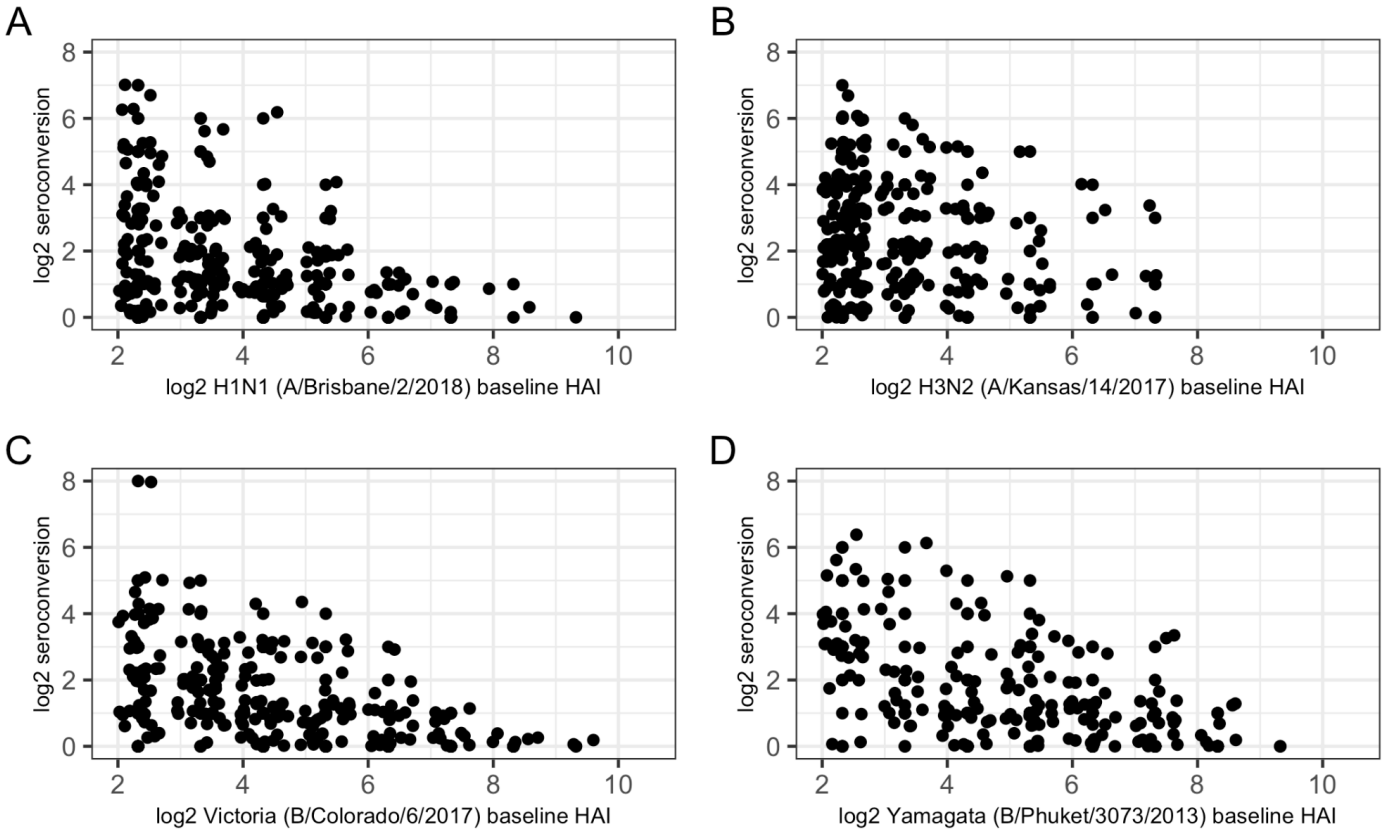


Figure S4. **Relationship between baseline HAI and seroconversion score for each strain.** The baseline HAI, measured prior to vaccination and the corresponding seroconversion score for the **A**) H1N1, **B**), H3N2, **C**) Victoria, and **D**) Yamagata strains included in the vaccine formulations. Values for all 275 individuals are reported for each of the three of the strains, but only the 214 individuals who received the standard dose formulation containing the Yamagata component are shown in panel D. A small amount of random noise was added to each value for visualization purposes to avoid overplotting points.

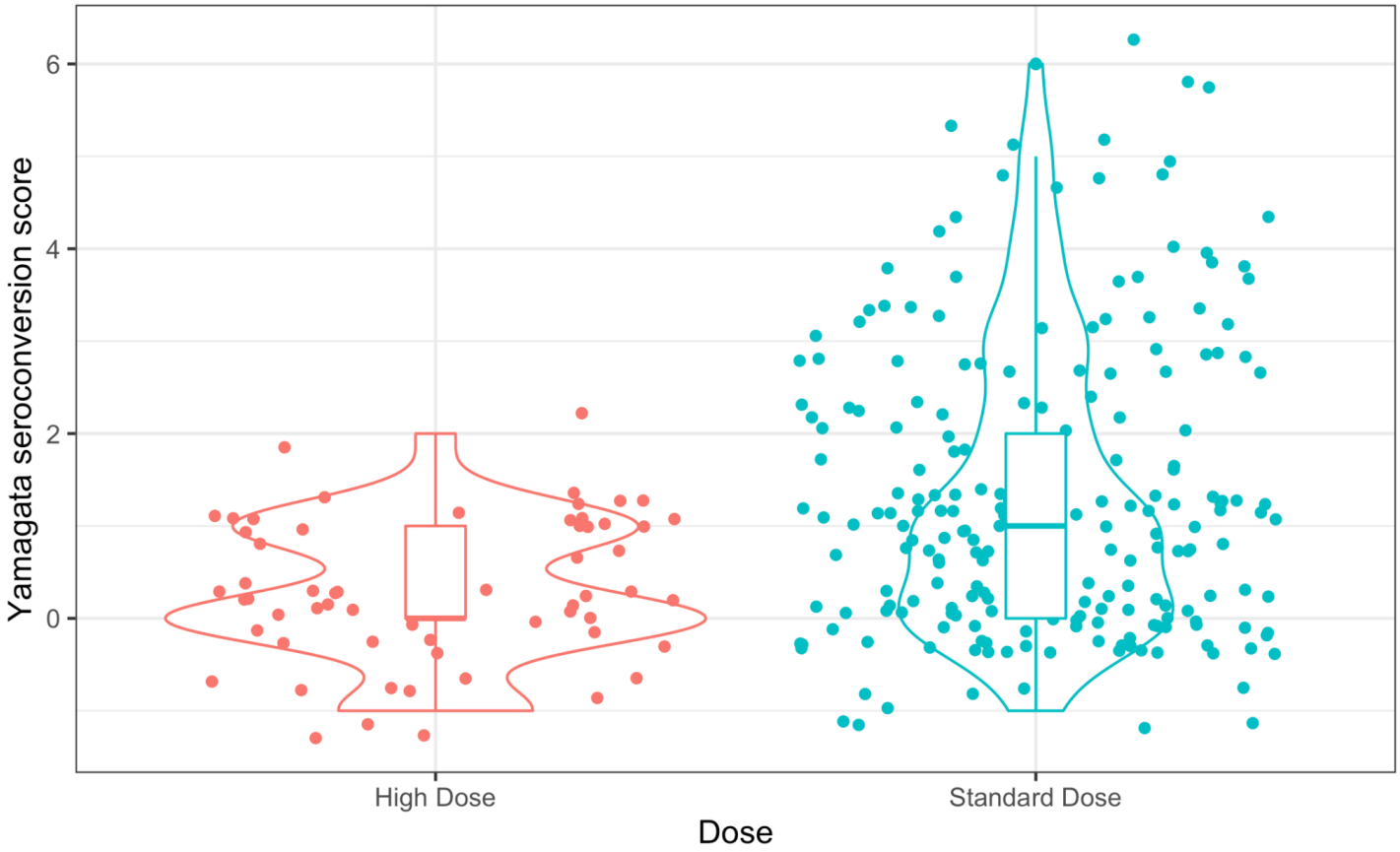


Figure S5. **Seroconversion differences between vaccine formulations.** The distribution of seroconversion scores for the Yamagata strain for those individuals (n = 61) who received the high dose formulation in which the Yamagata strain was excluded and those individuals (n = 214) who received the low dose formulation in which the Yamagata strain was included. A small amount of random noise was added to each value for visualization purposes to avoid overplotting points.

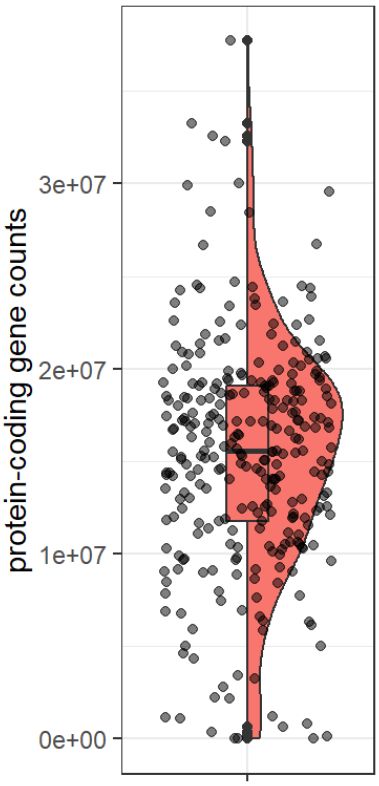


Figure S6. RNAseq read count distribution across 275 study participants. A median of 15 million sequence reads per individual mapped to protein coding genes.

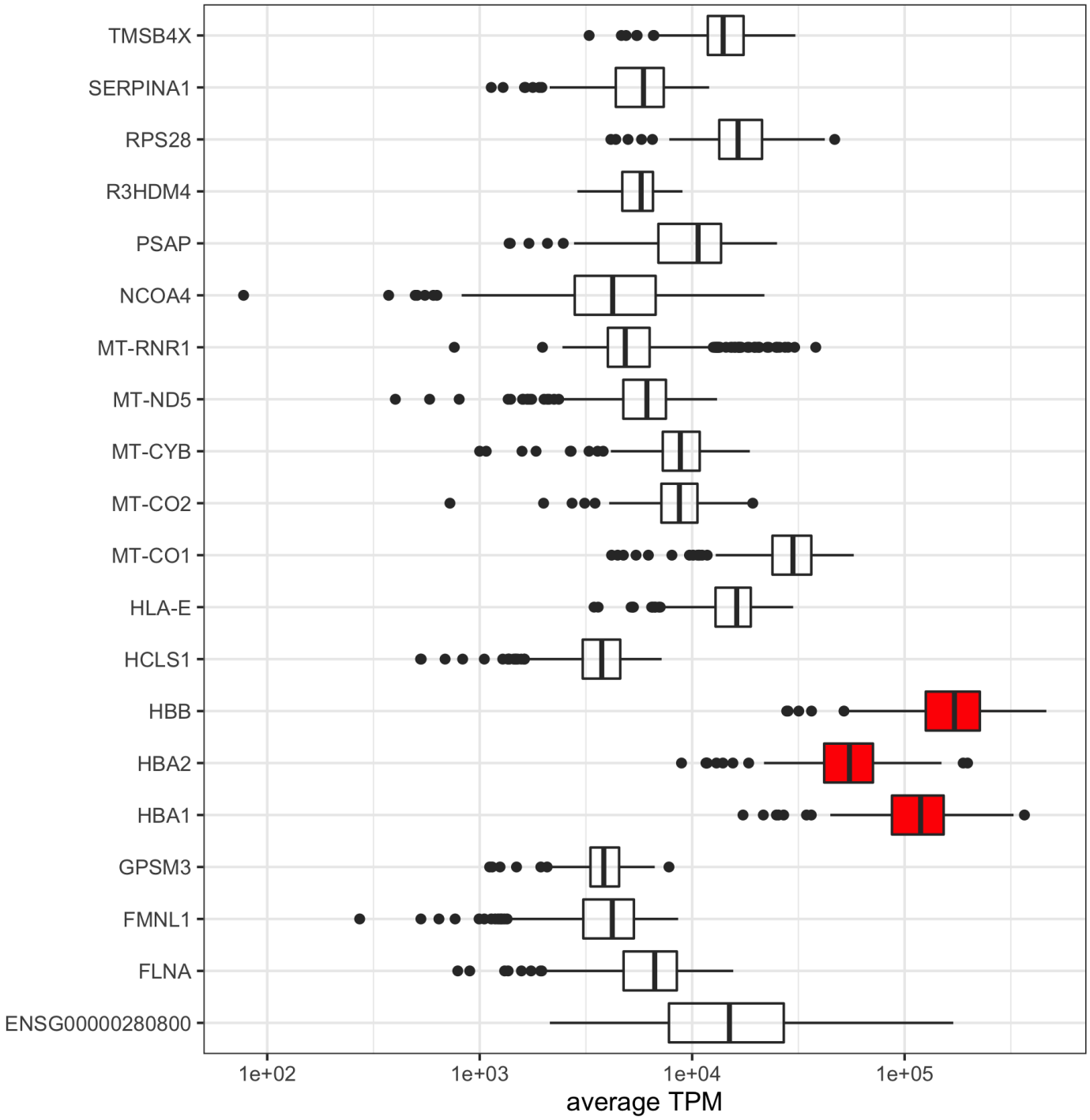


Figure S7. Genes with highest average transcripts per million (TPM) across all study participants. The aggregate counts of three genes, encoding hemoglobin, account for close to 50% of total transcripts reads.

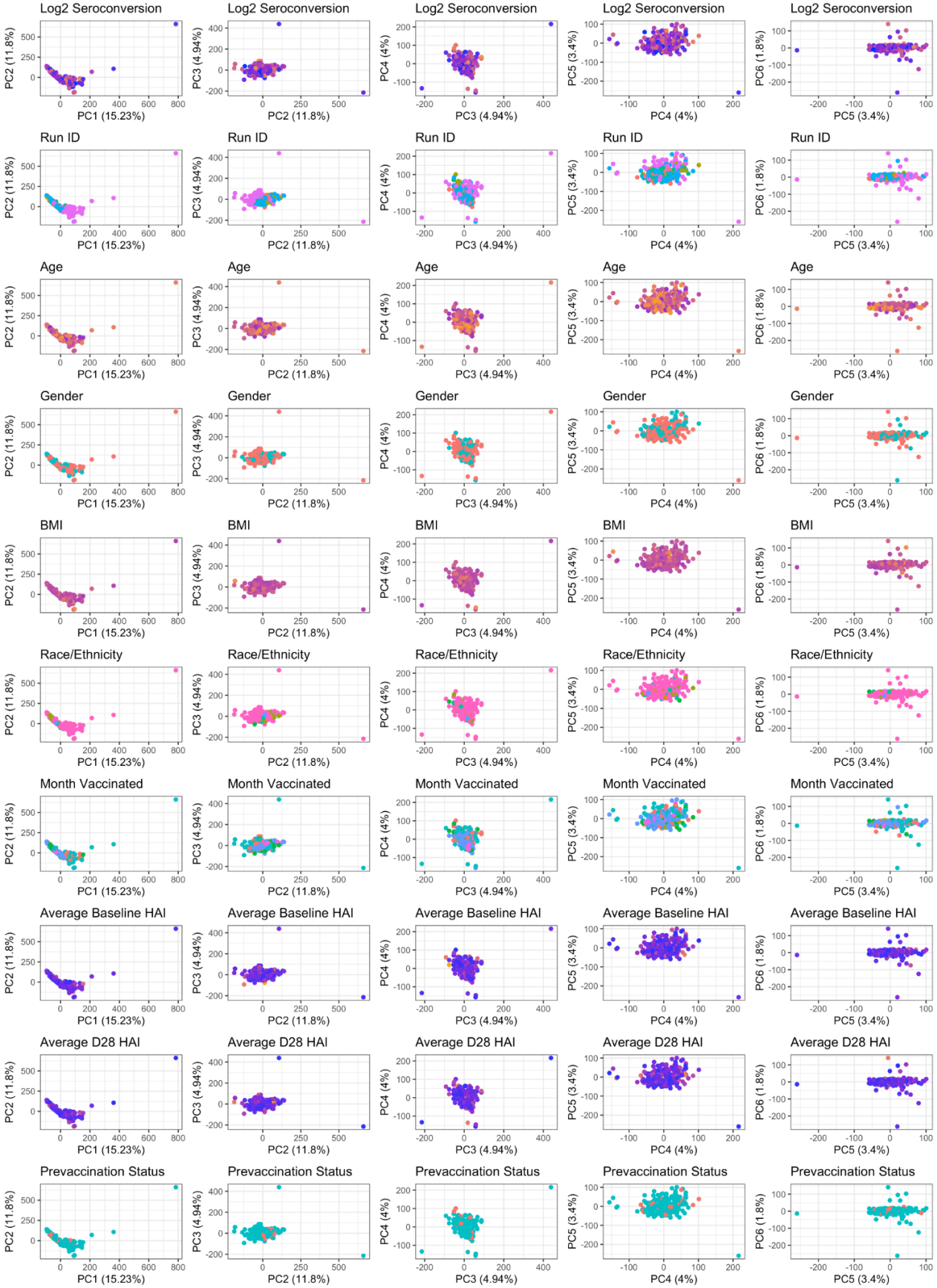


Figure S8. Principal component analysis of per sample transcript counts from RNA sequencing data.

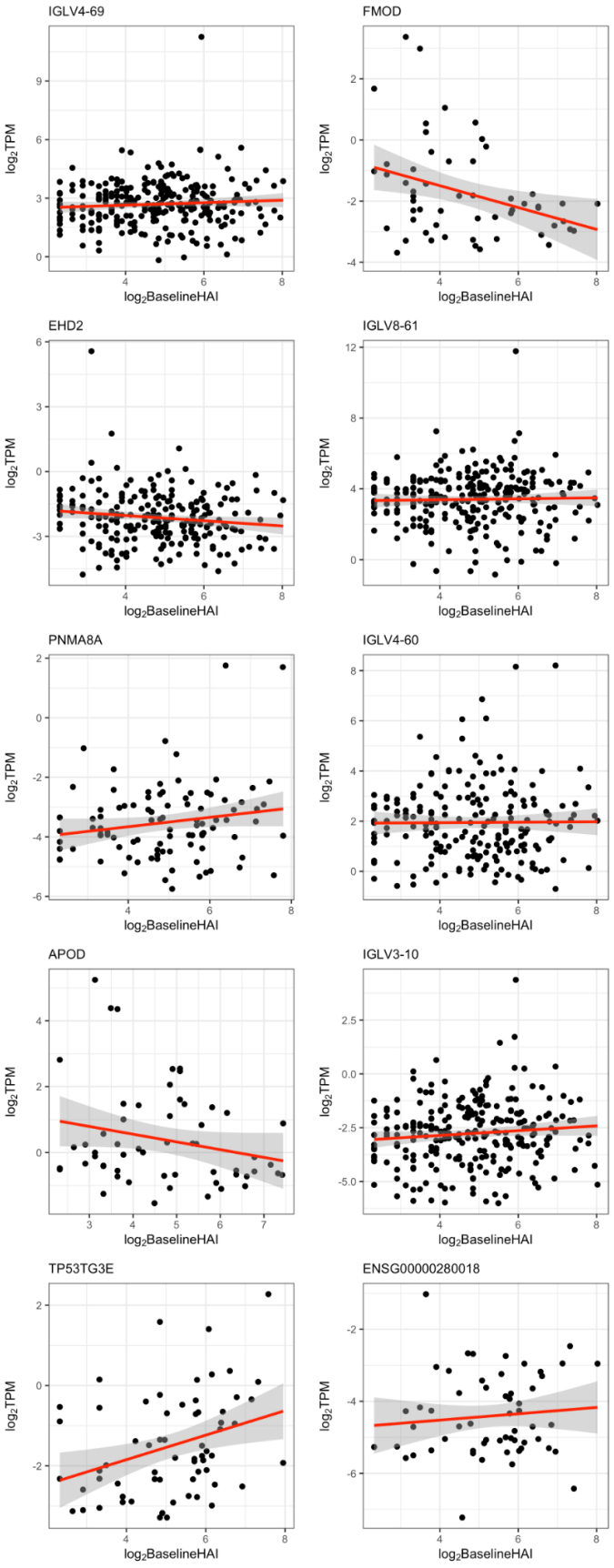


Figure S9. Top 10 DE genes as a function of baseline HAI ranked by FDR. Genes were selected for visualization on the basis of statistical significance and detection in at least 50 participants.

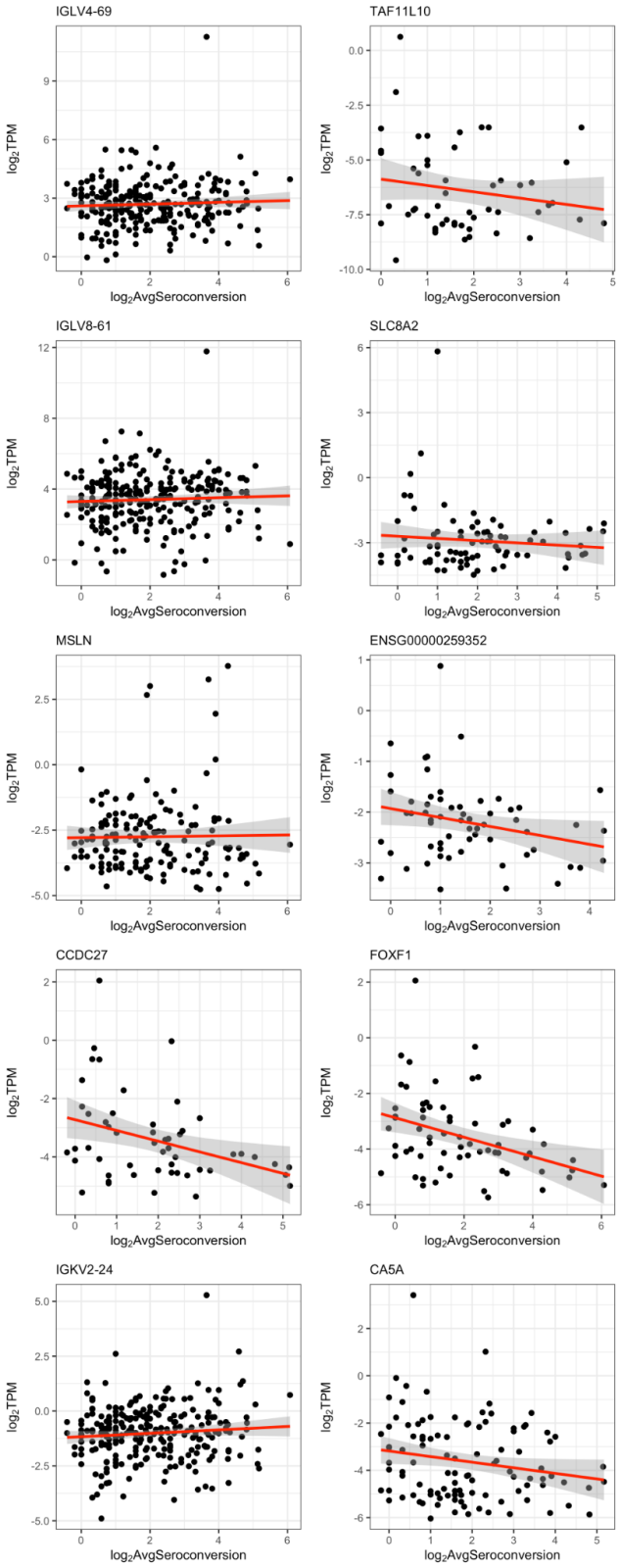


Figure S10. Top 10 DE genes as a function of seroconversion ranked by FDR. Genes were selected for visualization on the basis of statistical significance and detection in at least 50 participants.

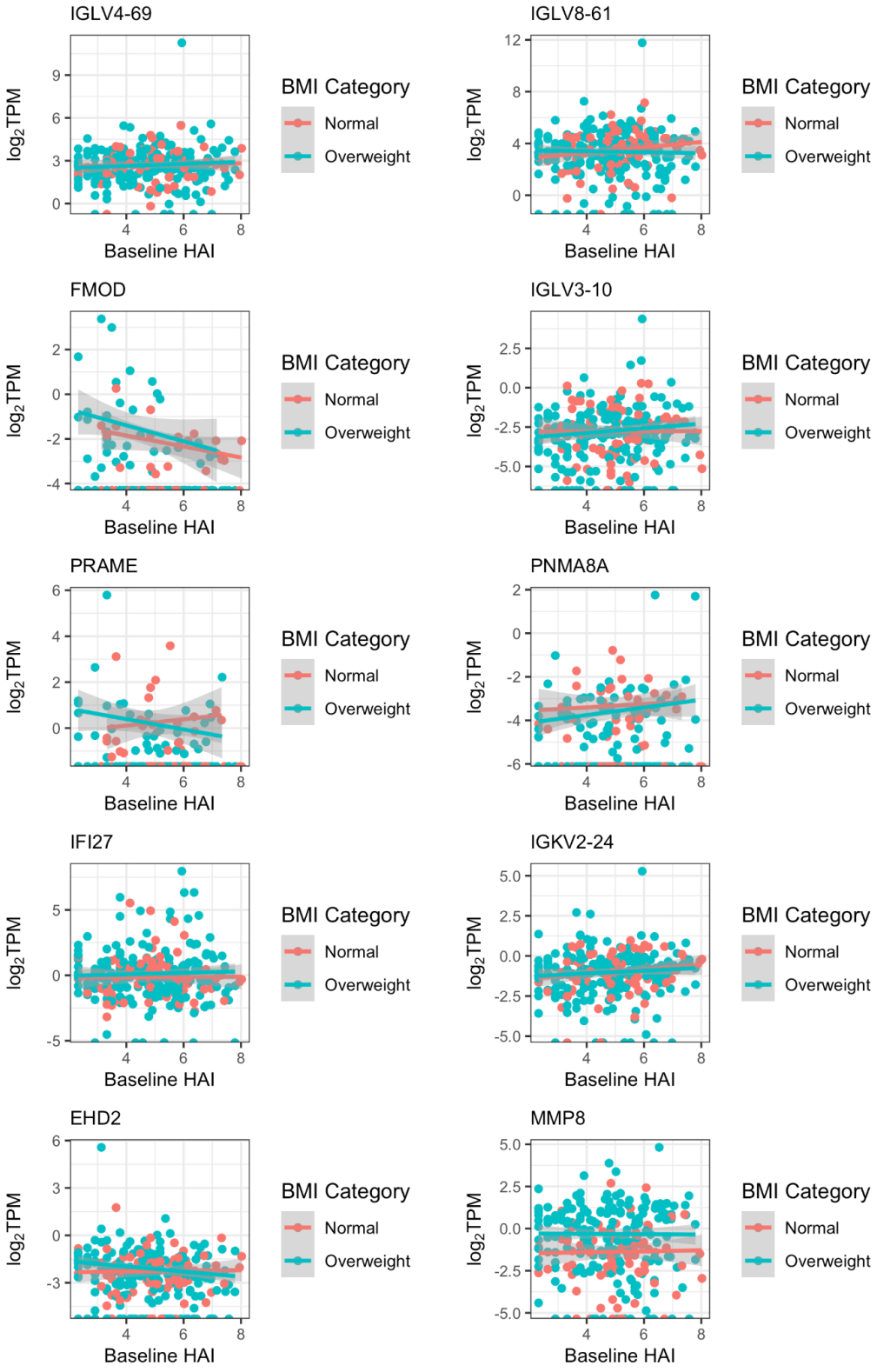


Figure S11. Top 10 DE genes as ranked by correlation coefficient to the interaction variable of initial HAI and BMI. Genes were selected for visualization on the basis of statistical significance and detection in at least 50 participants.

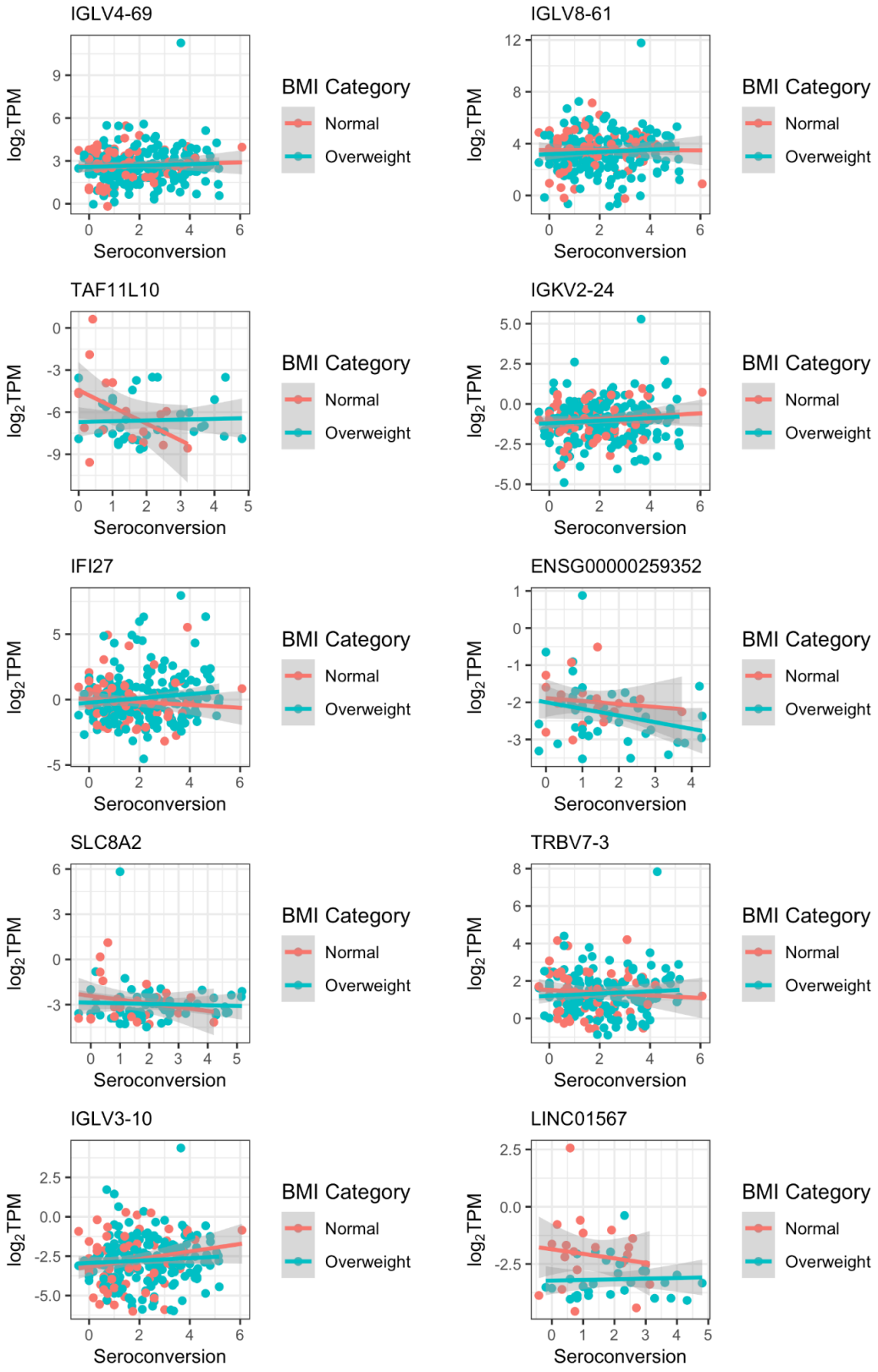


Figure S12. Top 10 DE genes as ranked by correlation coefficient to the interaction variable of seroconversion score and BMI. Genes were selected for visualization on the basis of statistical significance and detection in at least 50 participants.

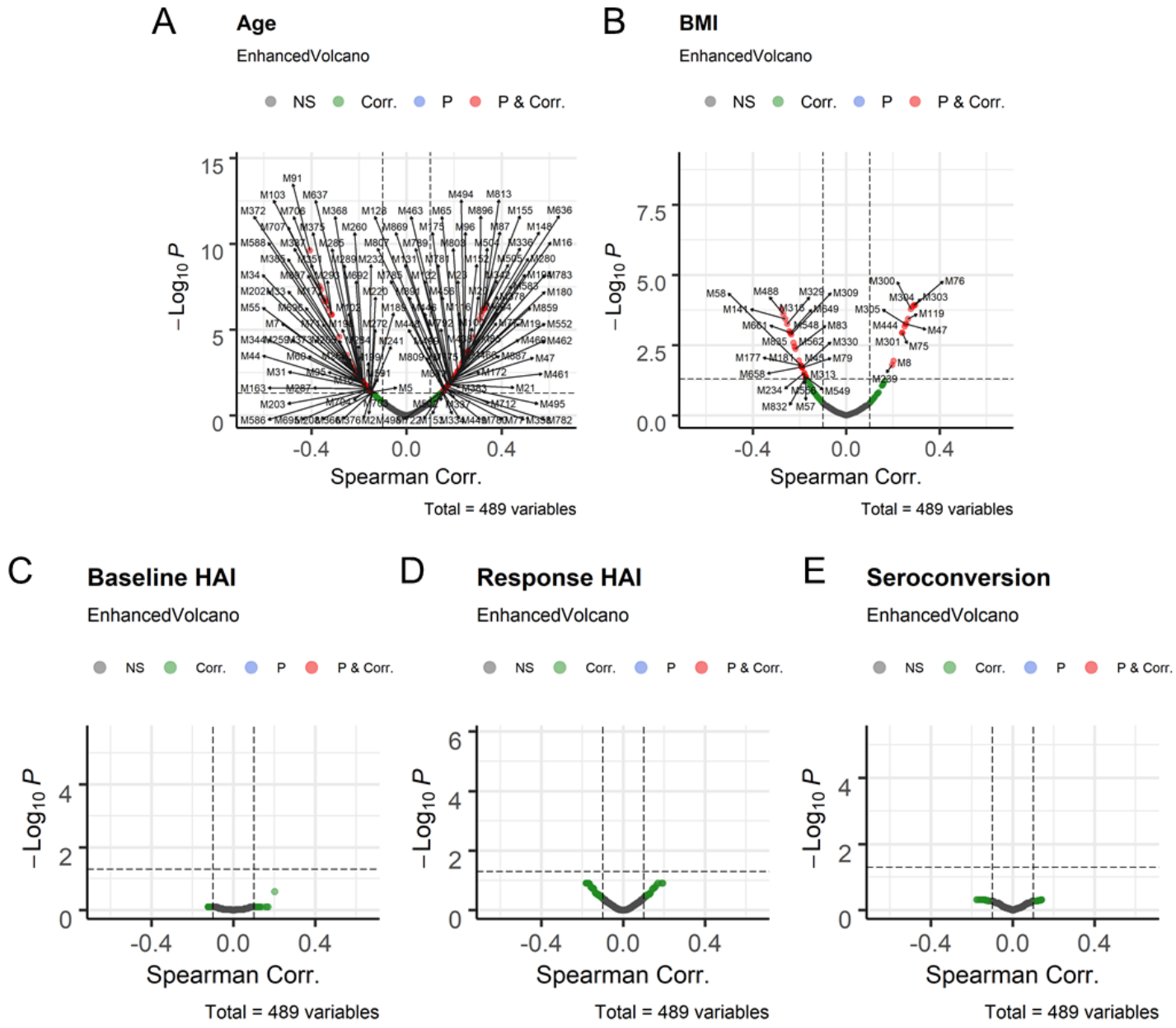
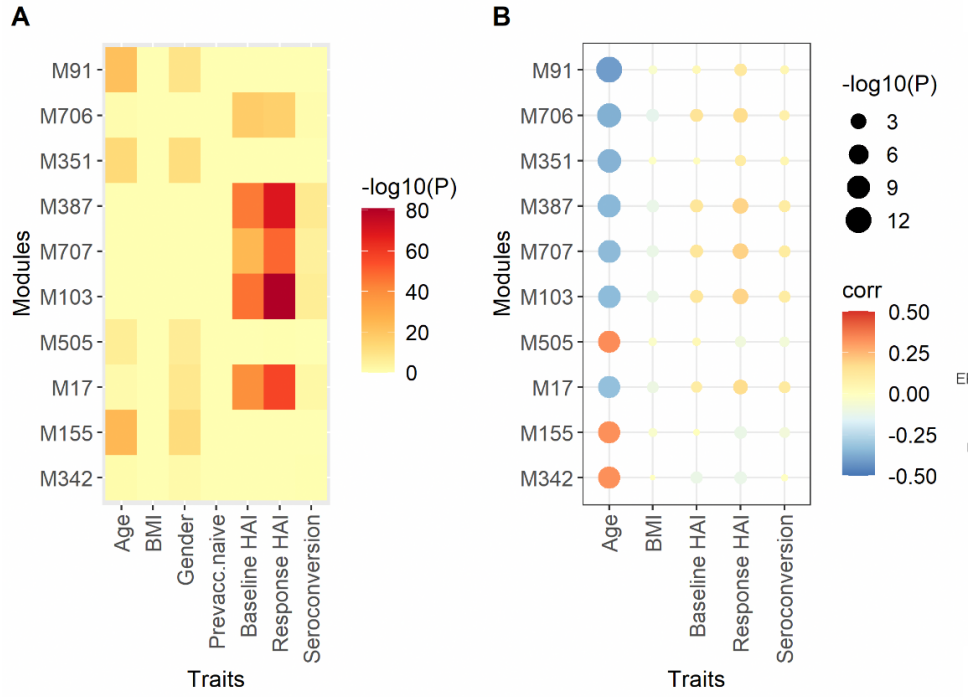
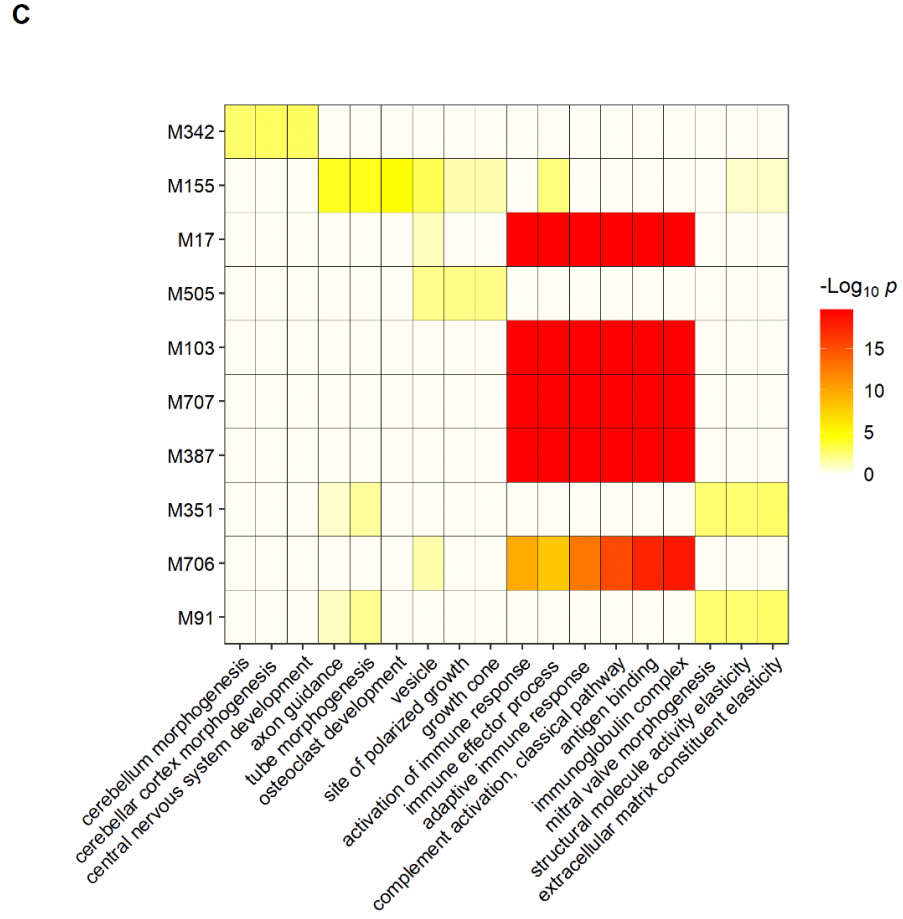


Figure S13. **Volcano Plots of Module / Trait relationships.**

Module/trait relationships are shown between the first principal component of MEGENA modules and selected quantitative traits. *P* on the y-axis refers to FDR (p-values adjusted for multiple testing).





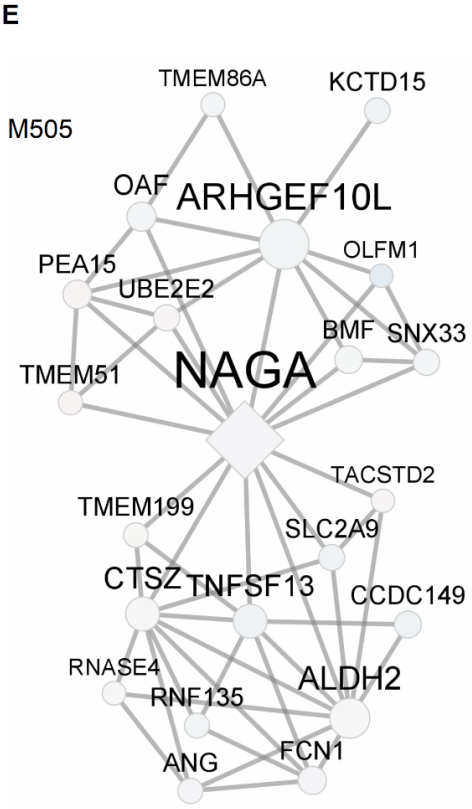


Figure S14. Best ranked modules after significant age response (related to Figure 4).

(A, B) The 10 best ranked modules are shown using a combined rank based on the correlations related to age. (A) The module enrichment for DEGs are depicted in this heatmap. (B) The dot-plot shows the correlation coefficients and corresponding P-value of module / trait correlations. (C) The functional enrichment of the corresponding modules for GO functions are shown. (D) The depicted module M91 is associated with the elasticity and recoil of the extracellular matrix. (E) The vesicle/lysosome component related module M505 is shown. Node colors in (D, E) indicate fold change with respect to high versus low average seroconversion, red for increased expression in subjects with high seroconversion compared to subjects with low seroconversion, blue refers to decreased gene expression between these two groups of subjects. The hierarchical relationship of the modules shown in (A) and (B) are as follows: M91 → M351; M155 → M505; and M17 → M103 → M387 → M706/M707 (see **Fig. 4**).

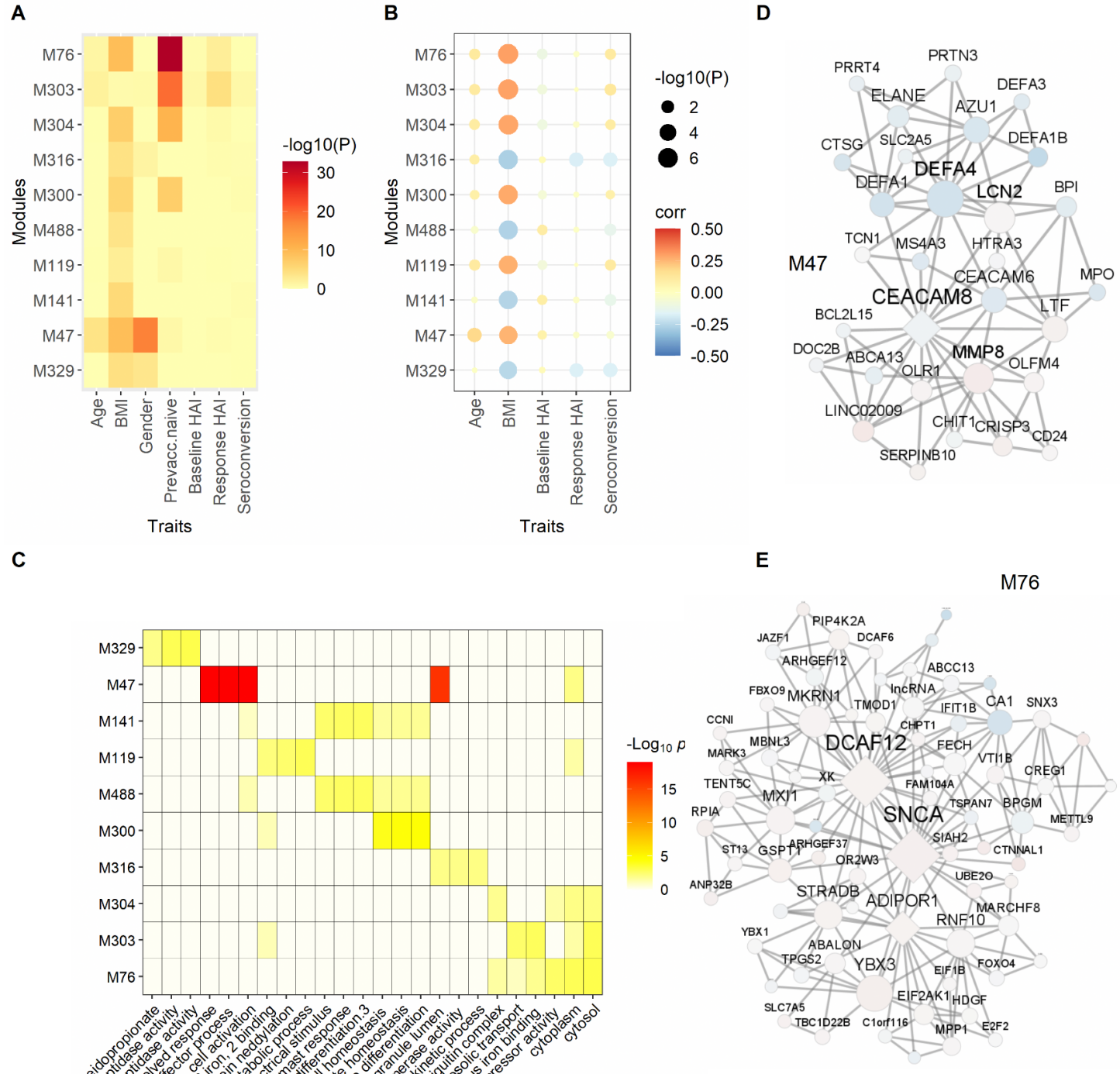
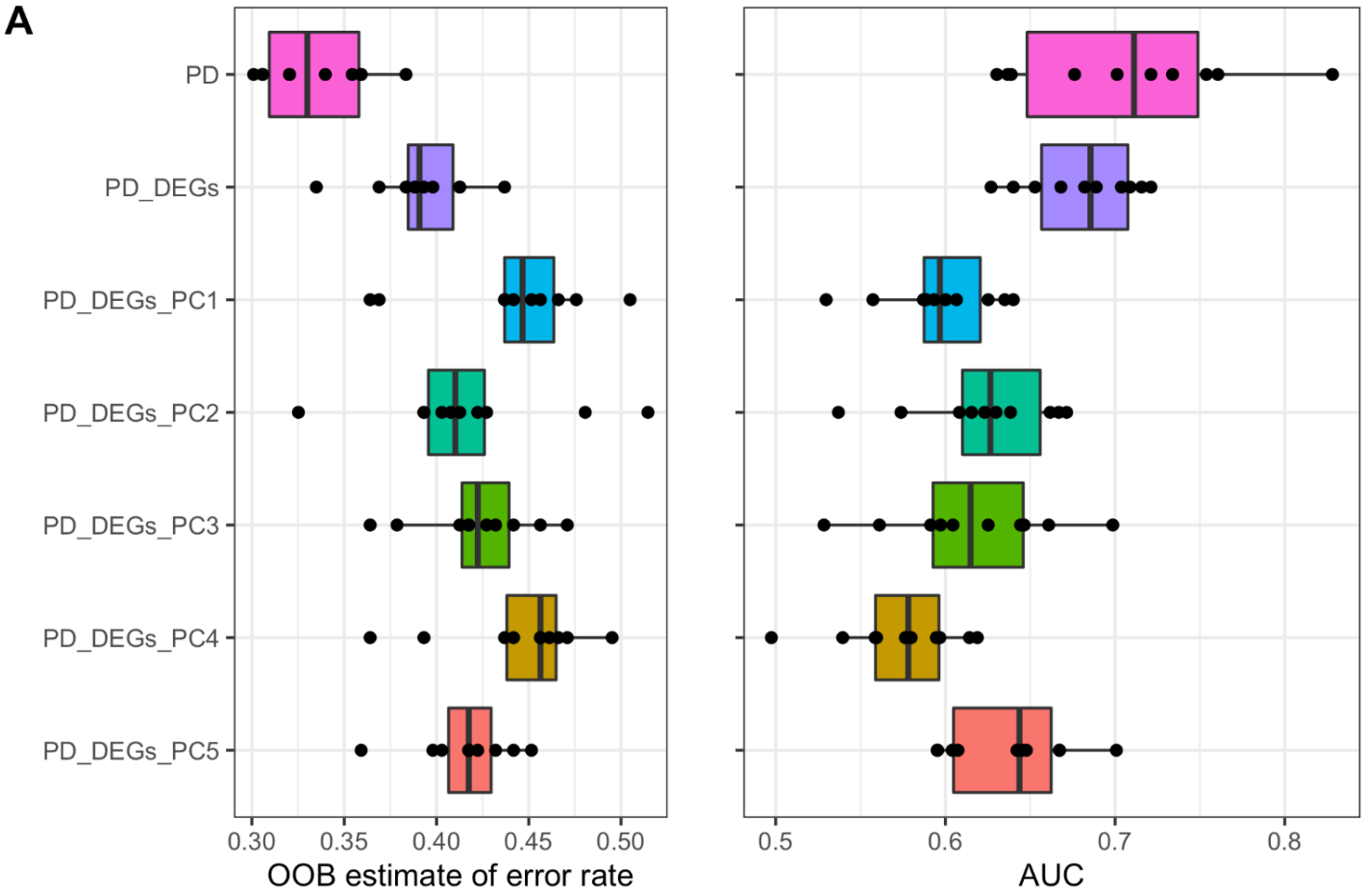


Figure S15. Best ranked modules after significant BMI response (related to Figure 4).

(A, B) The 10 best ranked modules are shown using a combined rank based on the correlations related to BMI. (A) The module enrichment for DEGs are depicted in this heatmap. (B) The dot-plot shows the correlation coefficients and corresponding P-value of module / trait correlations. (C) The functional enrichment of the corresponding modules for GO functions are shown. (D) Module M47 related to cell activation involved in immune response is shown. (E) Hemoglobin metabolism related module M8 is depicted. Node colors in (D, E) indicate fold change with respect to high versus low average seroconversion, red for increased expression in subjects with high seroconversion compared to subjects with low seroconversion, blue refers to decreased gene expression between these two groups of subjects. The hierarchical relationship of the modules shown in (A) and (B) are as follows: M76 → M303/M304; M141 → M488.



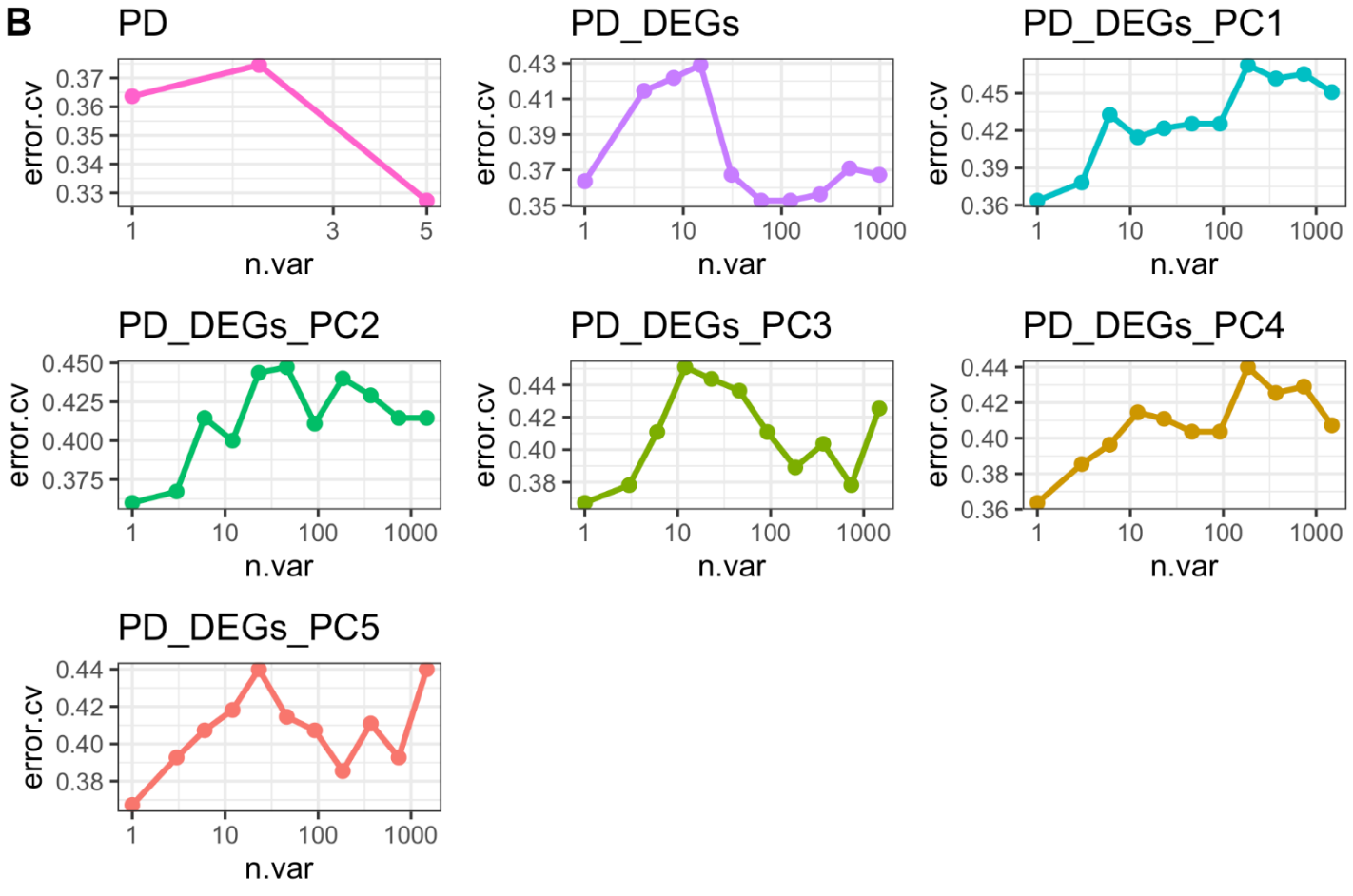


Figure S16. Random forest prediction performance of random forest models generated using clinical, single unit, and aggregated transcriptional data

The average **(A)** out-of-bag (OOB) error rate and **(B)** average area under the curve (AUC) values generated using 10 iterations of random forest models constructed using combinations of physiological data (PD), the 1,084 differentially expressed genes (DEGs) as a function of seroconversion score, and the principal components 1-5 of the aggregated module data. **(C)** A cross validation analysis plotting a varying number of variables for the clinical data (left), expression data (middle), and both combined (right) random forest models against the resulting cross validation error.

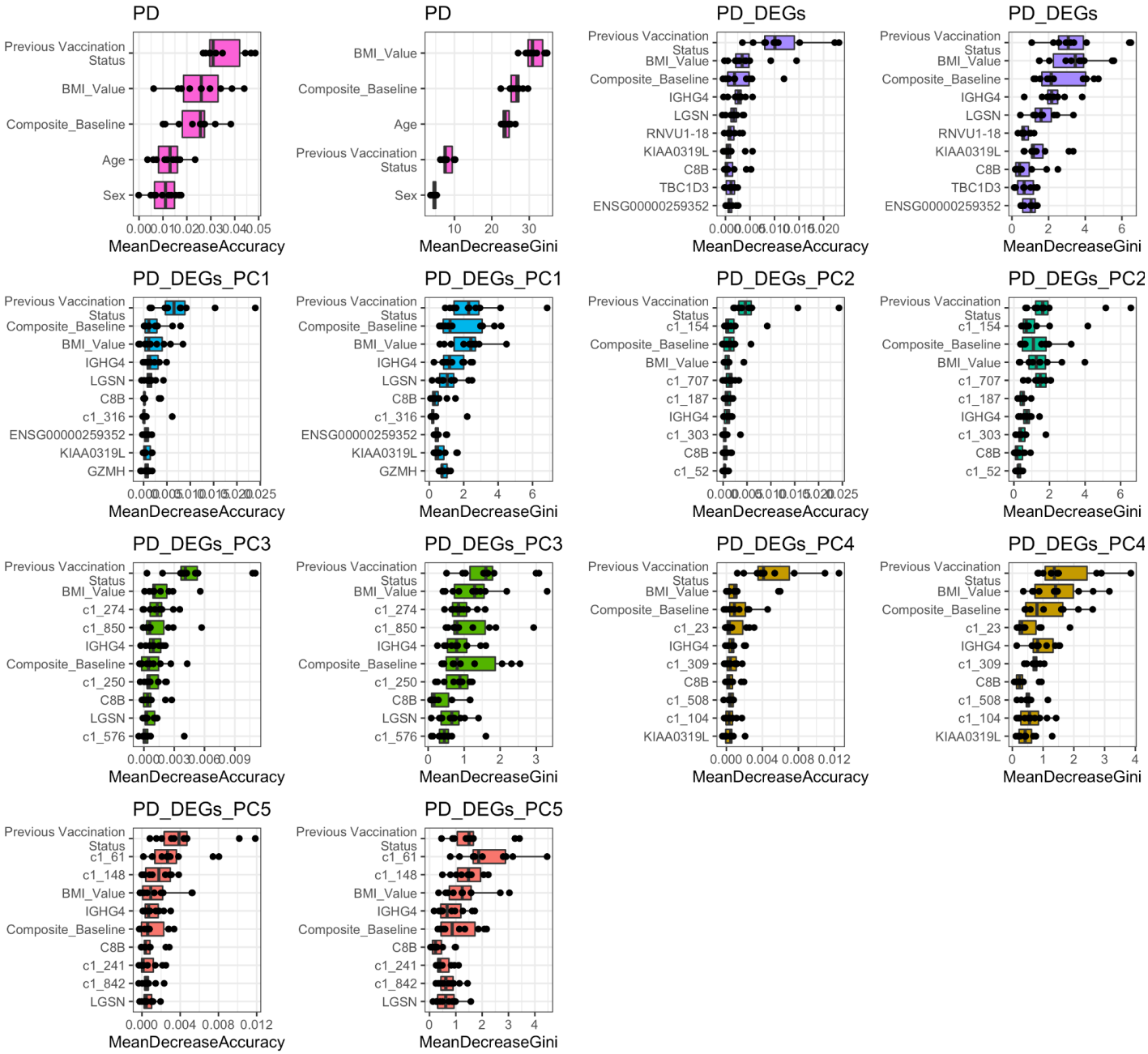


Figure S17. Variable importance values for top 10 clinical data, DEGs, and modules for random forest models generated with clinical data, DEGs, and module principal components

Variable importance measured through mean decrease accuracy and mean decrease in Gini coefficient for random forest modules generated using physiological data (PD), the 1,084 DEGs as a function of a seroconversion, and one of the five principal components of the expression module dataset.

Supplementary Tables

Table S1. UGA4 study participant information.

Table S2. Gene counts for each study participant obtained from RNAseq performed on whole blood.

Table S3. Univariate differential gene expression analysis of baseline HAI.

Table S4. Univariate differential gene expression analysis of seroconversion.

Table S5. Bivariate differential gene expression analysis of baseline HAI and BMI.

Table S6. Bivariate differential gene expression analysis of seroconversion and BMI.

References

Abreu, Rodrigo B., Greg A. Kirchenbaum, Emily F. Clutter, Giuseppe A. Sautto, and Ted M. Ross. 2020. "Preexisting Subtype Immunodominance Shapes Memory B Cell Recall Response to Influenza Vaccination." *JCI Insight* 5 (1). https://doi.org/10.1172/jci.insight.132155.

Adamson, Penelope J., Mahmood A. Al Kindi, Jing J. Wang, Alex D. Colella, Timothy K. Chataway, Nikolai Petrovsky, Tom P. Gordon, and David L. Gordon. 2017. "Proteomic Analysis of Influenza Haemagglutinin-Specific Antibodies Following Vaccination Reveals Convergent Immunoglobulin Variable Region Signatures." *Vaccine* 35 (42): 5576–80.

Aw, Danielle, Alberto B. Silva, and Donald B. Palmer. 2007. "Immunosenescence: Emerging Challenges for an Ageing Population." *Immunology* 120 (4): 435–46.

Bartholomeus, Esther, Nicolas De Neuter, Pieter Meysman, Arvid Suls, Nina Keersmaekers, George Elias, Hilde Jansens, et al. 2018. "Transcriptome Profiling in Blood before and after Hepatitis B Vaccination Shows Significant Differences in Gene Expression between Responders and Non-Responders." *Vaccine* 36 (42): 6282–89.

Blachly, James S., Amy S. Ruppert, Weiqiang Zhao, Susan Long, Joseph Flynn, Ian Flinn, Jeffrey Jones, et al. 2015. "Immunoglobulin Transcript Sequence and Somatic Hypermutation Computation from Unselected RNA-Seq Reads in Chronic Lymphocytic Leukemia." *Proceedings of the National Academy of Sciences of the United States of America* 112 (14): 4322–27.

CDC. 2021a. "Defining Adult Overweight & Obesity." Centers for Disease Control and Prevention. June 7, 2021. https://www.cdc.gov/obesity/adult/defining.html.

———. 2021b. "How Flu Vaccine Effectiveness and Efficacy Are Measured." August 31, 2021. https://www.cdc.gov/flu/vaccines-work/effectivenessqa.htm.

Chen, Yao-Qing, Teddy John Wohlbold, Nai-Ying Zheng, Min Huang, Yunping Huang, Karlynn E. Neu, Jiwon Lee, et al. 2018. "Influenza Infection in Humans Induces Broadly Cross-Reactive and Protective Neuraminidase-Reactive Antibodies." *Cell* 173 (2): 417–29.e10.

Cowling, Benjamin J., Wey Wen Lim, Ranawaka A. P. M. Perera, Vicky J. Fang, Gabriel M. Leung, J. S. Malik Peiris, and Eric J. Tchetgen Tchetgen. 2019. "Influenza Hemagglutination-Inhibition Antibody Titer as a Mediator of Vaccine-Induced Protection for Influenza B." *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 68 (10): 1713–17.

Cuenca-Zamora, Ernesto José, Francisca Ferrer-Marín, José Rivera, and Raúl Teruel-Montoya. 2019. "Tubulin in Platelets: When the Shape Matters." *International Journal of Molecular Sciences* 20 (14). https://doi.org/10.3390/ijms20143484.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1): 15–21.

Forgacs, David, Rodrigo B. Abreu, Giuseppe A. Sautto, Greg A. Kirchenbaum, Elliott Drabek, Kevin S. Williamson, Dongkyoon Kim, Daniel E. Emerling, and Ted M. Ross. 2021. "Convergent Antibody Evolution and Clonotype Expansion Following Influenza Virus Vaccination." *PloS One* 16 (2): e0247253.

Frasca, Daniela, Alain Diaz, Maria Romero, Denisse Garcia, and Bonnie B. Blomberg. 2020. "B Cell Immunosenescence." *Annual Review of Cell and Developmental Biology* 36 (October): 551–74.

Gaziano, Liam, Claudia Giambartolomei, Alexandre C. Pereira, Anna Gaulton, Daniel C. Posner, Sonja A. Swanson, Yuk-Lam Ho, et al. 2021. "Actionable Druggable Genome-Wide Mendelian Randomization Identifies Repurposing Opportunities for COVID-19." *Nature Medicine* 27 (4): 668–76.

Gene Ontology Consortium. 2015. "Gene Ontology Consortium: Going Forward." *Nucleic Acids Research* 43 (Database issue): D1049–56.

GSEA. Accessed June 11, 2022. https://www.gsea-msigdb.org/gsea/msigdb/genesets.jsp?collection=CP.

"Guidance for Industry Clinical Data Needed to Support the Licensure of Pandemic Influenza Vaccines." 2007. 2007. https://www.fda.gov/files/vaccines,%20blood%20&%20biologics/published/Guidance-for-Industry--Clinical-Data-Needed-to-Support-the-Licensure-of-Pandemic-Influenza-Vaccines.pdf.

HIPC-CHI Signatures Project Team, and HIPC-I Consortium. 2017. "Multicohort Analysis Reveals Baseline Transcriptional Predictors of Influenza Vaccination Responses." *Science Immunology* 2 (14). https://doi.org/10.1126/sciimmunol.aal4656.

Katz, Jacqueline M., Kathy Hancock, and Xiyan Xu. 2011. “Serologic Assays for Influenza Surveillance, Diagnosis and Vaccine Evaluation.” *Expert Review of Anti-Infective Therapy* 9 (6): 669–83.

Klein, Sabra L., and Andrew Pekosz. 2014. “Sex-Based Biology and the Rational Design of Influenza Vaccination Strategies.” *The Journal of Infectious Diseases* 209 Suppl 3 (July): S114–19.

Lex, Alexander, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. 2014. “UpSet: Visualization of Intersecting Sets.” *IEEE Transactions on Visualization and Computer Graphics* 20 (12): 1983–92.

Lidak, Tomas, Nikol Baloghova, Vladimir Korinek, Radislav Sedlacek, Jana Balounova, Petr Kasparek, and Lukas Cermak. 2021. “CRL4-DCAF12 Ubiquitin Ligase Controls MOV10 RNA Helicase during Spermatogenesis and T Cell Activation.” *International Journal of Molecular Sciences* 22 (10). https://doi.org/10.3390/ijms22105394.

Mandric, Igor, Jeremy Rotman, Harry Taegyun Yang, Nicolas Strauli, Dennis J. Montoya, William Van Der Wey, Jiem R. Ronas, et al. 2020. “Profiling Immunoglobulin Repertoires across Multiple Human Tissues Using RNA Sequencing.” *Nature Communications* 11 (1): 3126.

Marino, Stephen F., Uwe Jerke, Susanne Rolle, Oliver Daumke, and Ralph Kettritz. 2022. “Competitively Disrupting the Neutrophil-Specific Receptor-Autoantigen CD177:proteinase 3 Membrane Complex Reduces Anti-PR3 Antibody-Induced Neutrophil Activation.” *The Journal of Biological Chemistry* 298 (3): 101598.

Neidich, S. D., W. D. Green, J. Rebeles, E. A. Karlsson, S. Schultz-Cherry, T. L. Noah, S. Chakladar, M. G. Hudgens, S. S. Weir, and M. A. Beck. 2017. “Increased Risk of Influenza among Vaccinated Adults Who Are Obese.” *International Journal of Obesity* 41 (9): 1324–30.

Okeleji, Lateef Olabisi, Ayodeji Folorunsho Ajayi, Grace Adebayo-Gege, Victoria Oyetayo Aremu, Oluwadunsin Iyanuoluwa Adebayo, and Emmanuel Tayo Adebayo. 2021. “Epidemiologic Evidence Linking Oxidative Stress and Pulmonary Function in Healthy Populations.” *Chronic Diseases and Translational Medicine* 7 (2): 88–99.

Patel, Harshil, Phil Ewels, Alexander Peltzer, Rickard Hammarén, Olga Botvinnik, Gregor Sturm, Denis Moreno, et al. 2022. *Nf-Core/rnaseq: Nf-Core/rnaseq v3.7 - Iron Iguana*. https://doi.org/10.5281/zenodo.6513815.

Rima, Mohamad, Marwa Daghsni, Anaïs Lopez, Ziad Fajloun, Lydie Lefrancois, Mireia Dunach, Yasuo Mori, et al. 2017. “Down-Regulation of the Wnt/β-Catenin Signaling Pathway by Cacnb4.” *Molecular Biology of the Cell* 28 (25): 3699–3708.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. “edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.” *Bioinformatics* 26 (1): 139–40.

Robinson, Mark D., and Alicia Oshlack. 2010. “A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data.” *Genome Biology* 11 (3): R25.

Singhania, Akul, Joshua C. Wallington, Caroline G. Smith, Daniel Horowitz, Karl J. Staples, Peter H. Howarth, Stephan D. Gadola, Ratko Djukanović, Christopher H. Woelk, and Timothy S. C. Hinks. 2018. “Multitissue Transcriptomics Delineates the Diversity of Airway T Cell Functions in Asthma.” *American Journal of Respiratory Cell and Molecular Biology* 58 (2): 261–70.

Srivastava, Avi, Laraib Malik, Hirak Sarkar, Mohsen Zakeri, Fatemeh Almodaresi, Charlotte Soneson, Michael I. Love, Carl Kingsford, and Rob Patro. 2020. “Alignment and Mapping Methodology Influence Transcript Abundance Estimation.” *Genome Biology* 21 (1): 239.

Stacey, Hannah D., Neda Barjesteh, Jonathan P. Mapletoft, and Matthew S. Miller. 2018. “‘Gnothi Seauton’: Leveraging the Host Response to Improve Influenza Virus Vaccine Efficacy.” *Vaccines* 6 (2). https://doi.org/10.3390/vaccines6020023.

Stephenson, Iain, Rose Gaines Das, John M. Wood, and Jacqueline M. Katz. 2007. “Comparison of Neutralising Antibody Assays for Detection of Antibody to Influenza A/H3N2 Viruses: An International Collaborative Study.” *Vaccine* 25 (20): 4056–63.

Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, et al. 2005. “Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (43): 15545–50.

Thakar, Juilee, Subhasis Mohanty, A. Phillip West, Samit R. Joshi, Ikuyo Ueda, Jean Wilson, Hailong Meng, et al. 2015. “Aging-Dependent Alterations in Gene Expression and a Mitochondrial Signature of Responsiveness to Human Influenza Vaccination.” *Aging* 7 (1): 38–52.

WHO. 2018. “Influenza (seasonal).” November 6, 2018. https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal).

Wong, Guang W., Paul S. Foster, Shinsuke Yasuda, Jian C. Qi, Surendran Mahalingam, Elizabeth A. Mellor, Gregory Katsoulotos, et al. 2002. “Biochemical and Functional Characterization of Human Transmembrane Tryptase (TMT)/tryptase Gamma. TMT Is an Exocytosed Mast Cell Protease That Induces Airway Hyperresponsiveness in Vivo via an Interleukin-13/interleukin-4 Receptor Alpha/signal Transducer and Activator of Transcription (STAT) 6-Dependent Pathway.” *The Journal of Biological Chemistry* 277 (44): 41906–15.

Yu, Guangchuang, Li-Gen Wang, Yanyan Han, and Qing-Yu He. 2012. “clusterProfiler: An R Package for Comparing Biological Themes among Gene Clusters.” *Omics: A Journal of Integrative Biology* 16 (5): 284–87.