The Immune Signatures Data Resource: A compendium of systems vaccinology datasets

Joann Diray-Arce[1,2,*], Helen E.R. Miller[2,*], Evan Henrich[2,*], Bram Gerritsen[3], Matthew P. Mulè[4,5], Slim Fourati[6], Jeremy Gygi[3], Thomas Hagan[7], Lewis Tomalin[8], Dmitry Rychkov[9], Dmitri Kazmin[10], Daniel G. Chawla[3], Hailong Meng[3], Patrick Dunn[11], John Campbell[11], The Human Immunology Project Consortium (HIPC)[&], Minnie Sarwal[9], John S. Tsang[4], Ofer Levy[1,2,12], Bali Pulendran[7], Rafick Sekaly[6], Aris Floratos[13], Raphael Gottardo[2,14], Steven H. Kleinstein[3,**], Mayte Suárez-Fariñas[8,**]

*These authors contributed equally
**These authors contributed equally

Affiliations:
[1]Precision Vaccines Program, Boston Children's Hospital, [2]Harvard Medical School, Boston, MA, USA; [2]Fred Hutchinson Cancer Research Center, Seattle, WA, USA; [3]Yale School of Medicine, New Haven, CT, USA; [4]Multiscale Systems Biology Section, Laboratory of Immune System Biology, NIAID
NIH Center for Human Immunology, NIH, Bethesda, MD, USA; [5]NIH-Oxford-Cambridge Scholars Program, Department of Medicine, Cambridge University; [6]Emory University School of Medicine, Atlanta, GA, USA; [7]Stanford University School of Medicine, Stanford University, Stanford, CA, USA; [8]Icahn School of Medicine at Mount Sinai, New York, New York, USA; [9]University of California, San Francisco, San Francisco, CA, USA; [10]The Jackson Laboratory for Genomic Medicine, Farmington CT; [11]*ImmPort* Curation Team, NG Health Solutions, Rockville, MD, USA; [12]Broad Institute of MIT & Harvard, Cambridge, MA, USA. [13]Columbia University Medical Center, New York, NY, USA [14]University of Lausanne and University Hospital of Lausanne, Lausanne, Switzerland

Abstract:

Vaccines are among the most cost-effective public health interventions for preventing infection-induced morbidity and mortality, yet much remains to be learned regarding the mechanisms by which vaccines protect. Systems immunology combines traditional immunology with modern 'omic profiling techniques and computational modeling to promote rapid and transformative advances in vaccinology and vaccine discovery. The NIH/NIAID Human Immunology Project Consortium (HIPC) has leveraged systems immunology approaches to identify molecular signatures associated with the immunogenicity of many vaccines, including those targeting seasonal influenza, yellow fever, and hepatitis B. These data are made available to the broader scientific community through the *ImmuneSpace* data portal and analysis engine leveraging the NIH/NIAID *ImmPort* repository[1,2]. However, a barrier to progress in this area is that comparative analyses have been limited by the distributed nature of some data, potential batch effects across studies, and the absence of multiple relevant studies from non-HIPC groups in *ImmPort*. To support comparative analyses across different vaccines, we have created the Immune Signatures Data Resource, a compendium of standardized systems vaccinology datasets. This data resource is available through *ImmuneSpace*, along with code to reproduce the processing and batch normalization starting from the underlying study data in *ImmPort* and the Gene Expression Omnibus (GEO). The current release comprises 1405 participants from 53 cohorts profiling the response to 24 different vaccines and includes transcriptional profiles and antibody response measurements. This novel systems vaccinology data release represents a valuable resource for comparative and meta-analyses that will accelerate our understanding of mechanisms underlying vaccine responses.

Design Type(s)	☒ Factorial design ☒ Longitudinal design ☒ Transcription profiling design
Factor Type(s)	☒ Sample type (PBMC, Blood) ☒ Vaccine ☒ Pathogen ☒ Age groups
Measurement Type(s)	☒ Transcription profiling assay ☒ Neutralizing antibody assay ☒ ELISA ☒ Influenza hemagglutination inhibition assay (HAI)
Technology Type(s)	☒ Microarray ☒ RNA sequencing
Sample Characteristic(s)	☒ Homo sapiens

Background and Summary

Vaccines, one of humanity's greatest public health achievements, save millions of lives every year by preventing infectious diseases[3,4]. Despite their widespread use and efficacy, much remains to be learned regarding their molecular mechanisms of action. This is true both for vaccines against pandemic infections such as influenza[5], and SARS-coronavirus-2[6], as well as for infections for which there are currently no authorized or approved vaccines such as HIV[7-9]. Elucidating the commonalities and differences in the immune responses induced by different vaccines and their association with protective antibody responses will provide deeper insight and a framework for the evidence-based design of better vaccines or vaccination strategies. Recent technologies have provided tools to probe the immune response to vaccination and integrate hierarchical levels of the biological system[10]. Alluded to as systems vaccinology[11], this new application of systems biology tools provides new insights into molecular mechanisms of vaccine-induced immunogenicity and protection[12-15].

The National Institute of Allergy and Infectious Diseases (NIAID) established a multi-institutional consortium, Human Immunology Project Consortium (HIPC)[2,16] , to characterize the immune system in diverse populations in response to a stimulus, such as vaccination, using high-dimensional 'omic platforms and modern computational tools[2]. Since the inception of the consortium in 2010, members of HIPC have published > 500 articles, including many that describe molecular signatures associated with vaccine-induced protection. These studies include molecular signatures that predict the immunogenicity of vaccination against yellow fever[17-20], seasonal influenza in healthy young adults, elderly[21-25], and children[26], shingles[27,28], dengue[29,30], malaria[31,32], and meta-analyses of common signatures across different vaccines[33,34]. These molecular signatures resulted from large-scale data analysis using high-throughput systems biology approaches coupled with detailed clinical phenotyping in well-characterized human cohorts.

Predicting immunogenicity from 'omic signatures remains challenging, prompting methodological innovation to advance the field towards delivering on the promises of precision vaccination[35-37]. The factors that contribute to robust vaccination responses are highly complex and span multiple biological scales. The vast collection of high-dimensional profiling datasets poses significant challenges for comparative analysis of these studies, including biological variability as well as data challenges such as volume, technical noise, and diverse sample processing pipelines. Data integration of cellular and molecular signatures to predict vaccine responses requires harmonization and normalization of data from multiple sources[38]. The generation of big data poses simultaneous challenges and opportunities with the potential of contributing to precision medicine. The biological interpretation of the resulting molecular features correlated with robust responses is another key factor. Understanding how effective vaccines stimulate protective immune responses, and how these mechanisms may differ between

vaccine types and targeted pathogens remains a substantial challenge for the field. Moreover, the systems vaccinology field has been limited by a lack of a formal framework to standardize immune signatures gathered from diverse studies, creating a bottleneck for comparative analysis. To address these challenges, and in support of advances in systems vaccinology by the HIPC project and the broader scientific community, we present the creation of the Immune Signatures Data Resource, a compendium of systems vaccinology studies that enables standardized comparative analysis to identify molecular signatures that correlate with the outcomes of vaccinations.

The current release of the Immune Signatures Data Resource consists of 4795 transcriptomic samples from 1405 participants curated from 30 *ImmPort* studies (16 from HIPC-related studies, 14 non-HIPC studies) (Figure 2). The transcriptomic profiling dataset is derived from 53 cohorts of 820 young adults (18-49 years old) and 585 (≥50 years old) older adult samples. The data resource covers 24 vaccines targeting 11 pathogens and 6 vaccine types (Figure 1B, 4A), thus creating a critical mass of data that will serve as a valuable resource for the broader scientific community. Additionally, data assembly and integration of these data set enables derivation of comparable signatures for each study for comparative analysis of the underlying data.

Methods:

Database background information and structure

Compatibility with ImmPort and ImmuneSpace, the central databases of the Human Immunology Project Consortium: Given the exponential growth of the number of datasets of multiple modalities, an urgent need emerged for data sharing across the broader scientific community. The HIPC implements the NIH Data Sharing policy to promote the principles of Findability, Accessibility, Interoperability, and Reusability (FAIR) via *ImmPort*, created under the National Institute of Allergy and Infectious Diseases Division of Allergy, Immunology, and Transplantation (NIAID-DAIT). *ImmPort* (*ImmPort*.org) is an open repository of participant-level large-scale human immunology data designed to aid scientists with data standards and guidelines for efficient secondary analyses[1,39]. *ImmPort* facilitates data sharing of immunology studies creating a centralized knowledge base and resources, and serves as a central data repository for HIPC. *ImmuneSpace*[2,34] extends *ImmPort*, providing access to additional data (e.g., standardized gene expression matrices) and web-based R tools for data accession, analysis, and reporting. Studies in the Immune Signatures Data Resource are archived through the Shared Data Portal on *ImmPort* and *ImmuneSpace* repositories and may be updated over time. To provide a consistent data source for reproducible results, we also archived a static copy of the data as a "virtual study" in *ImmuneSpace* (Figure 1A & 2).

Identification of vaccine study cohorts with transcriptomic profiles: Through a literature search conducted from 2017 to 2020, we identified target publications with systems-level profiling of human vaccination responses. We found 16 HIPC-funded vaccinology studies in *ImmPort* with transcriptomics datasets generated with matching immune response outcomes. Notably, we have supplemented the HIPC data previously available in *ImmPort* by curating and submitting 14 additional human vaccination studies to *ImmPort*. For studies that were not in *ImmPort*/*ImmuneSpace*, we located the underlying data by surveying public transcriptome databases (e.g., Gene Expression Omnibus (GEO)) or reaching out to study authors to request data access, allowing us to submit to *ImmPort* on their behalf. These datasets were then made available via *ImmuneSpace* to be processed for standardization, preprocessing checks, and normalization. The standard analytical pipeline enables reproducibility and comparability of future studies to be correlated with publicly available immune response measurement. This process created the virtual study for the HIPC named the Immune Signatures Data Resource (Figure 1A, Figure 2).

Gene Expression Data processing pipeline

Data were read directly from *ImmuneSpace* using *ImmuneSpaceR* functions and subsequently preprocessed, quality controlled, and integrated using the following pipeline:

Quality Control of Microarray experiments: The *ArrayQualityMetrics* R package[40] was used for quality control and assurance of all microarray experiments (Figure 3A). Outlier detection was based on the following statistics: i) Mean absolute difference of M-values (log-ratios) of each pair of arrays, ii) the Kolmogorov-Smirnov statistic K_a between each array's signal intensity distribution and the distribution of the pooled data and, iii) the Hoeffding's statistic D_a on the joint distribution of A (average) and M values for each array. Using pre-specified criteria within an established public microarray data reuse pipeline[40], we flagged for removal arrays that failed all three quality control statistics.

Preprocessing: Raw probe intensity data for Affymetrix studies were background-corrected and summarized using the RMA algorithm[41] while the function *read.ilmn* (*limma* R package) was used to read and background correct Illumina raw probe intensities. To integrate RNA-seq and microarray data, raw counts for RNA-seq data were converted to log-transformed values incorporating observational level weights to account for technical variations using the *voom*[42] transformation. Expression data within each study were quantile normalized and log-transformed separately for each cohort/sample type.

Annotation*:* We annotated the manufacturing IDs (probes from microarray/Illumina) to their corresponding gene alias. Gene aliases were mapped to the recent gene symbols from the HUGO Gene Nomenclature Committee[43] [accessed Dec 23, 2020]. For the rare case where a gene alias mapped to more than one gene symbol, the mapping was resolved by the following: i) If a gene alias mapped to itself as a symbol, as well as other symbols, then it was mapped to itself; ii) if the gene alias mapped to multiple

symbols that did not include itself, then the gene alias was dropped from the study. As a result, the raw gene expression matrix was reduced to 10086 HUGO gene aliases with known unique mapping.

Gene-based expression profiles: Expression data were summarized at the probe level (for microarray data) and gene-alias level (RNA-seq) to the canonical Gene-Symbol level. The probes / gene-aliases were summarized by selecting the probe or gene-alias with the highest average expression (mean of probes across all samples, take the highest mean) across all samples within the matrix (cohort and sample type).

Cross-Study normalization*:* One of the main assumptions in expression analysis is that differences in gene expression across conditions occur in a relatively small number of processes. As such, the distribution across conditions should be similar, and departures of these assumptions are corrected, for example, using quantile normalization. This procedure usually creates a target distribution using all samples available, but we observed dissimilar distributions in our collection stemming from various platforms used. Such differences lead to extensive distributions and introduce artifacts in the data (Figure 3B and 3C). The target distribution was obtained from samples using Affymetrix platforms, resulting in a well-defined distribution, and each sample in our collection was quantile normalized to this target distribution. Before cross-study normalization, there were 35,725 representative gene symbols present. There were 25,639 genes removed after normalization, as these genes were not present in all the studies. This yielded a final expression matrix of 4795 samples from 1405 participants representing 10,086 genes (Figure 2).

Determining and adjusting for technical confounders: We studied the primary sources of variation in the data, including the study effect (which also encompasses the impact of different expression platforms (RNA-seq, Affymetrix arrays, Illumina arrays, etc.), sample types (Whole blood, PBMC), as well as demographics. We conducted Principal Component Analysis (PCA) to visualize such associations in a bidimensional space of principal components (PCs) and applied Principal Variance Component Analysis (PVCA)[44] to quantify the amount of variability attributed to different experimental conditions. This approach models the multivariate distribution of the PCs computed for the PCA as a function of experimental factors and estimates the total variance explained by each factor via mixed-effect models. Since many studies included only one vaccine, temporal variations due to vaccine response were confounded with the study effect. The assessment of the primary technical sources of variation was carried out using only the pre-vaccination data, not affected by the targeted pathogen and vaccine type used in the different studies. Of note, all studies enrolled healthy volunteers, and the first biosample was obtained pre-vaccination. The targeted pathogen and vaccine type should not affect these baseline data.

Platform, study, and sample types were identified as significant sources of variation in the gene expression matrix. The effect of those three variables was estimated by modeling gene expression at baseline (at which no vaccine or timepoint effect exists) with a linear model using the *limma* framework,

including feature set vendor (Platform/Affy), study (batch factors), and sample type, Y-chromosome genes presence, as covariates. Study and cell-type effects were estimated using a linear model with age, Y-chromosome genes presence (biological sex), study, sample type (Whole Blood/PBMC), study, and platform as additive effects. From here, the study, platform, and cell-type effects were eliminated from the entirety of the expression matrix. There were three studies (SDY1276, SDY1264, SDY180) that contained multiple cohorts and were treated as separate studies.

***Biological sex imputation*:** Imputation of biological sex, as defined by the presence of a Y-chromosome, was carried out based on the gene expression profiles of 13 Y-chromosome genes. Within each study, a multidimensional scaling was first applied to the Y-chromosome gene expression profiles. K-means clustering was then used to cluster samples into two groups. Participants in the cluster with higher mean expression values were considered male (i.e., the Y-chromosome was present) while those in the cluster with lower expression were considered female (i.e., the Y-chromosome was absent). The consistency of the Y-chromosome presence assignment across time points was verified (Figure 3D). In the (few) cases where imputation was not in agreement across all time points, the reported sex was used and if no sex was reported, imputation followed a majority rule principle.

Age Imputation: Age imputation for studies without reported ages (SDY1260, SDY1264, SDY1293, SDY1294, SDY1364, SDY1370, SDY1373, SDY984) employed the RAPToR R v1.1.5 package[45]. The RAPToR algorithm takes in a reference set of gene expression time series with reported ages and generates a near-continuous, high-temporal resolution from the interpolated reference dataset. Transcriptomic profiles of participants without reported ages were compared to the reference dataset via a correlation profile, providing age estimates for the sample. Finally, random subsets of genes from the subject's transcriptomic profile were bootstrapped to ascertain a confidence interval for the imputed age. We generated the reference dataset using the transcriptomic profiles of 21 studies in our resource for which age was reported. The studies were split into younger (age < 50) and older (age ≥ 50) cohorts, thus two different models were generated, and only baseline transcriptomic profiles were used in the reference dataset. As RAPToR also enables phenotypic data to be incorporated into the interpolation model, each possible combination of phenotypic features was tested. For each combination, RAPToR predicted the age for participants in the 21 studies with known age, and the goodness of fit was evaluated by the coefficient of determination (R^2). The best model for the younger and older cohorts was then used to impute ages for the 7 studies without reported age (Figure 3E, 3F)

Immune response datasets processing pipeline:

To identify the molecular signatures that correlate with vaccine immunogenicity, we included immune response readouts in the creation of this data resource. For studies that were missing vaccine response endpoints in their public data deposition, we contacted study authors and requested available antibody

response measures to vaccine antigens. Once shared, these data were submitted to *ImmPort* and linked to the relevant studies. These readouts include neutralizing antibody titers (Nab), hemagglutination inhibition assay (HAI) results for influenza studies, and Immunoglobulin IgG ELISA assay results. In participants for whom the humoral immune response was measured with multiple assays, the preference was given to HAI for influenza or Nab for non-influenza studies, then IgG ELISA datasets. The antibody measures were normalized within each study by estimating the fold-change differences between the post-vaccination time-point (generally between day 28 or day 30) compared to the baseline measurement. For influenza studies where the vaccine included multiple strains, the fold changes between the post-vaccination versus baseline were calculated for each strain, and the maximum fold change (MFC) over the strains was selected[34]. Due to the variability in baseline antibody (Ab) levels and immune memory such as influenza vaccines, we also estimated the maximum residual after baseline adjustment (maxRBA) method by calculating the maximum residual across all vaccine strains to adjust for variable baseline Ab levels using the R package *titer*[21]. A total of 30 studies with 1405 participants and 4795 samples have both transcriptomics and immune response readout data available (Figure 2). This dataset enables researchers to carry out comparative analyses using immunogenicity data as well as prediction of the quality of response across multiple vaccines.

Data Records:

The Immune Signatures Data Resource is available online for download by the research community from this website: The data is hosted on *ImmuneSpace* and can be accessed via the R package *ImmuneSpaceR* (https://rglab.github.io/ImmuneSpaceR/). The resource is available for use by the scientific community and can be downloaded from a research data repository IS2 https://www.*ImmuneSpace*.org/is2.url. A summary of datasets, with their corresponding study ID and accession numbers, is provided in Table 4.

Technical Validation

Quality Control and Assurance

For global quality control across all public microarray data, we used a well-established pipeline available through the *ArrayQualitymetrics* R package[40]. Using pre-specified criteria established in the existing public microarray data reuse pipeline[46], arrays that failed 3 out of 3 calculated quality control statistics were flagged for removal (see Methods). Consistent with standard practice to perform such quality control analysis prior to downstream analysis and dataset submission to the Gene Expression Omnibus, none of the samples were outliers by all three statistics (Fig 3A). As expected for data from published peer-reviewed studies, all the identified studies passed the quality assurance method using the *Arrayqualitymetrics* method.

Y-chromosomal presence and age imputation

A few studies were missing information for sex and for age. To achieve data completeness, we included the biological sex imputation based on the imputed presence of the Y-chromosome using gene expression, as well as imputation of age when the variable was missing or defined by a broad range of values. Age imputation employed the RAPToR tool using 21 studies with reported age to define the best predictive model for the younger (age < 50 years) and older (age ≥ 50 years) cohorts separately. The highest correlation coefficients from the young cohort were generated by taking into account the model (X ~ age_reported + matrix) with a correlation coefficient of R^2=0.367 (Figure 3E), while the old cohort yielded a prediction R^2 of 0.536 for their highest coefficient value (Figure 3F).

Definition of Vaccination Studies Transcriptomic Cohort

Data preprocessing in *ImmuneSpace* yielded a total of 30 studies and 59 cohorts, with 1482 participants and 5413 samples. After the data was preprocessed and quality control measures were performed, we further assessed the identified cohorts as defined in the flow diagram (Figure 2). This curation included: i) removing participants that were not relevant to the objective (n=34); ii) removing samples due to inconsistencies with time design determination (n= 178); iii) removing participants with no baseline expression data (n=42). Some studies, such as SDY1368 and SDY67, were dropped from the normalized data sets as they did not include subjects within our target age range (18-50 years). In summary, we report that the final Immune Signatures Data Resource contains 53 cohorts from 30 studies with 1405 participants and 4795 samples.

Assessment and adjustment of the batch effects

We evaluated the main sources of variation on the gene expression matrix to identify and adjust technical confounders (RNA-seq, Affymetrix arrays, Illumina arrays, etc.), study, and specimen types (e.g., whole blood vs. PBMCs) using the baseline samples. Since all studies enrolled healthy volunteers, and the first sample was taken pre-vaccination, pathogen and vaccine type would not affect the baseline data. Figure 3B clearly demonstrates robust clustering of samples by study, which are also grouped by platform type. The study effect and type of platform used accounted for the vast majority (95%) of variation, followed by specimen types (3.6%). It is thus essential that the data are corrected for these major effects prior to any analytical usage [see Materials and Methods for further details]. The study, platform type, and specimen type-specific effects were estimated using a linear model that also included age and Y-chromosome presence as additive effects using only baseline expression. Once the study, platform, and specimen-type effects were estimated, they were eliminated from the entirety of the expression matrix. Figure 3B shows that those effects can successfully be adjusted from the data, thus leading to a matrix of expression that is free of most technical biases induced by the laboratory and cell-type effects.

Immune Signatures Transcriptomics and Immune Response Datasets

We report the total number of assay samples collected from the transcriptomic and immune response datasets tallied by targeted pathogen and vaccine type, across multiple systems vaccinology datasets (Figure 4A). We captured about ~3000 HAI antibody titer results from influenza studies that were measured by the standard HAI assay pre- and at multiple time points post-vaccination, depending on the study. Mean titers were calculated for the reported strains of the virus and were based on the highest dilution reported at day 28-30 post-vaccination. In addition, neutralizing antibody (NAB) titers and IgG ELISA results specific to each pathogen were determined by each study and are summarized (Figure 4A). The overall transcriptomics dataset comprises multiple time points from 7 days pre-vaccination up to day 180 days post-vaccination (Figure 4B). While most of the datasets focus on the young adult population (ages 18-50 years old), the data resource also includes studies that profile older adults following hepatitis B, influenza, and varicella vaccination (Figure 4C) that may be useful for analysis. The Euler diagram describes the dataset overlap of participants with transcriptomics datasets and corresponding to one or more immune response datasets (Figure 4D).

Heterogeneity of the immune response to vaccination across targeted pathogens and vaccine types was reflected in variation in the longitudinal trajectories of HAI and NAB titer measurements (Figure 5A and 5B). HAI and NAB titers generally increased by 14-28 days after vaccination but attenuated at different times for each vaccine (Figure 5A and 5B). Change in NAB titers after vaccination were significantly different across the 5 unique combinations of targeted pathogen and vaccine types where these measurements were reported (ANOVA $p < 10^{-10}$), with significant differences across all 5 groups except between meningococcus and yellow fever vaccines (Figure 5C). Some influenza vaccination studies reported both HAI and NAB measures of immunogenicity, and there was a significant positive correlation between the vaccination-induced changes in these titers across participants (Spearman's rho = 0.45, $p < 10^{-10}$) (Figure 5D).

Usage Notes

The expression data and accompanying meta-data have been made available with different formats and options to ease usage. Data are available as standard expression sets (eSet) objects, the R/Bioconductor structure unifying expression values, metadata, and gene annotation. Both normalized data and batch-adjusted data are available (Table 4). Users interested in a single study or those planning to work exclusively within participants' changes may opt for the normalized data without batch adjustment. For comparison of time points across studies or developing algorithms that use expression data, batch corrected matrices should be employed. Imputed age values for participants with no reported age were included to facilitate the use of age as a covariate in future analysis. Such analysis can be carried out with the complete data set and can be followed up by a sensitivity analysis using the small cohort with age-

reported data. For the use of expression sets with the corresponding immune response per participant, these are available in eSets noted with a response. The selected immune response outcome per study is also summarized in Table 3.

Code Availability

The source codes for the Immune Signatures Data Resource and all data are available in ImmuneSpace (https://www.immunespace.org/is2.url). Pre-processing code and supplementary data can be found in the ImmuneSignatures2 R package hosted on Github (https://github.com/RGLab/ImmuneSignatures2).

Acknowledgments

This research was conducted within the Human Immunology Project Consortium (HIPC) and supported by the National Institute of Allergy and Infectious Diseases. This work was supported in part by NIH grants U19AI128949, U19AI118608, U19AI118626, and U19AI089992, U19AI090023, U19AI089992, U19AI128914, U19AI118610, U19AI128913. The HIPC projects are listed at https://www.immuneprofiling.org/hipc/page/showPage?pg=projects. This work was supported in part by the Canadian Institutes of Health Research [funding reference number FDN-154287]

Contributions

All authors identified the datasets, performed quality control, and assurance and analyzed the datasets. JD-A, HM, SHK and MSF led the writing and organization of the manuscript. HM, EH, PD, together with the *ImmuneSpace* and *ImmPort* team, implemented the pipeline for data access and visualization. The HIPC Consortium contributed to the conception and design of the work, as well as the acquisition of data. A full list of the HIPC Consortium members can be found in Supplementary File 1. All authors edited and approved the manuscript.

Corresponding authors

Please address correspondence to joann.arce@childrens.harvard.edu or mayte.suarezfarinas@mssm.edu.

Ethics declaration

S.H.K. receives consulting fees from Northrop Grumman and Peraton. OL is an inventor on several patents relating to vaccine adjuvants and human *in vitro* systems predicting vaccine action. R.G. has received consulting income from Illumina, Takeda, and declares ownership in Ozette Technologies and Modulus Therapeutics. The other authors declare no competing interests.

REFERENCES:

1 Bhattacharya, S. *et al.* ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Sci Data* **5**, 180015, doi:10.1038/sdata.2018.15 (2018).

2 Brusic, V., Gottardo, R., Kleinstein, S. H., Davis, M. M. & committee, H. s. Computational resources for high-dimensional immune analysis from the Human Immunology Project Consortium. *Nat Biotechnol* **32**, 146-148, doi:10.1038/nbt.2777 (2014).

3 Piot, P. *et al.* Immunization: vital progress, unfinished agenda. *Nature* **575**, 119-129, doi:10.1038/s41586-019-1656-7 (2019).

4 Pulendran, B. Systems vaccinology: probing humanity's diverse immune systems with vaccines. *Proc Natl Acad Sci U S A* **111**, 12300-12306, doi:10.1073/pnas.1400476111 (2014).

5 Fineberg, H. V. Pandemic preparedness and response--lessons from the H1N1 influenza of 2009. *N Engl J Med* **370**, 1335-1342, doi:10.1056/NEJMra1208802 (2014).

6 Fauci, A. S., Lane, H. C. & Redfield, R. R. Covid-19 - Navigating the Uncharted. *N Engl J Med*, doi:10.1056/NEJMe2002387 (2020).

7 Fauci, A. S. An HIV Vaccine Is Essential for Ending the HIV/AIDS Pandemic. *JAMA* **318**, 1535-1536, doi:10.1001/jama.2017.13505 (2017).

8 Fauci, A. S., Folkers, G. K. & Marston, H. D. Ending the global HIV/AIDS pandemic: the critical role of an HIV vaccine. *Clin Infect Dis* **59 Suppl 2**, S80-84, doi:10.1093/cid/ciu420 (2014).

9 Fauci, A. S. & Marston, H. D. Ending the HIV-AIDS Pandemic--Follow the Science. *N Engl J Med* **373**, 2197-2199, doi:10.1056/NEJMp1502020 (2015).

10 Diercks, A. & Aderem, A. Systems approaches to dissecting immunity. *Curr Top Microbiol Immunol* **363**, 1-19, doi:10.1007/82_2012_246 (2013).

11 Pulendran, B., Li, S. & Nakaya, H. I. Systems Vaccinology. *Immunity* **33**, 516-529, doi:http://dx.doi.org/10.1016/j.immuni.2010.10.006 (2010).

12 Tsang, J. S. *et al.* Improving Vaccine-Induced Immunity: Can Baseline Predict Outcome? *Trends Immunol* **41**, 457-465, doi:10.1016/j.it.2020.04.001 (2020).

13 Nakaya, H. I., Li, S. & Pulendran, B. Systems vaccinology: learning to compute the behavior of vaccine induced immunity. *Wiley Interdiscip Rev Syst Biol Med* **4**, 193-205, doi:10.1002/wsbm.163 (2012).

14 Nakaya, H. I. & Pulendran, B. Systems vaccinology: its promise and challenge for HIV vaccine development. *Curr Opin HIV AIDS* **7**, 24-31, doi:10.1097/COH.0b013e32834dc37b (2012).

15 Zak, D. E. & Aderem, A. Overcoming limitations in the systems vaccinology approach: a pathway for accelerated HIV vaccine development. *Curr Opin HIV AIDS* **7**, 58-63, doi:10.1097/COH.0b013e32834ddd31 (2012).

16 Poland, G. A., Quill, H. & Togias, A. Understanding the human immune system in the 21st century: the Human Immunology Project Consortium. *Vaccine* **31**, 2911-2912, doi:10.1016/j.vaccine.2013.04.043 (2013).

17 Muyanja, E. *et al.* Immune activation alters cellular and humoral responses to yellow fever 17D vaccine. *J Clin Invest* **124**, 3147-3158, doi:10.1172/JCI75429 (2014).

18 Gaucher, D. *et al.* Yellow fever vaccine induces integrated multilineage and polyfunctional immune responses. *J Exp Med* **205**, 3119-3131, doi:10.1084/jem.20082292 (2008).

19 Querec, T. D. *et al.* Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. *Nat Immunol* **10**, 116-125, doi:10.1038/ni.1688 (2009).

20 Querec, T. *et al.* Yellow fever vaccine YF-17D activates multiple dendritic cell subsets via TLR2, 7, 8, and 9 to stimulate polyvalent immunity. *J Exp Med* **203**, 413-424, doi:10.1084/jem.20051720 (2006).

21 Avey, S. *et al.* Seasonal Variability and Shared Molecular Signatures of Inactivated Influenza Vaccination in Young and Older Adults. *J Immunol* **204**, 1661-1673, doi:10.4049/jimmunol.1900922 (2020).

22 Nakaya, Helder I. *et al.* Systems Analysis of Immunity to Influenza Vaccination across Multiple Years and in Diverse Populations Reveals Shared Molecular Signatures. *Immunity* **43**, 1186-1198, doi:http://dx.doi.org/10.1016/j.immuni.2015.11.012 (2015).

23 Nakaya, H. I. *et al.* Systems biology of vaccination for seasonal influenza in humans. *Nat Immunol* **12**, 786-795, doi:10.1038/ni.2067 (2011).

24 Oh, Jason Z. *et al.* TLR5-Mediated Sensing of Gut Microbiota Is Necessary for Antibody Responses to Seasonal Influenza Vaccination. *Immunity* **41**, 478-492, doi:http://dx.doi.org/10.1016/j.immuni.2014.08.009 (2014).

25 Thakar, J. *et al.* Aging-dependent alterations in gene expression and a mitochondrial signature of responsiveness to human influenza vaccination. *Aging (Albany NY)* **7**, 38-52, doi:10.18632/aging.100720 (2015).

26 Nakaya, H. I. *et al.* Systems biology of immunity to MF59-adjuvanted versus nonadjuvanted trivalent seasonal influenza vaccines in early childhood. *Proc Natl Acad Sci U S A* **113**, 1853-1858, doi:10.1073/pnas.1519690113 (2016).

27 Li, S. *et al.* Metabolic Phenotypes of Response to Vaccination in Humans. *Cell* **169**, 862-877 e817, doi:10.1016/j.cell.2017.04.026 (2017).

28 Sullivan, N. L. *et al.* Breadth and Functionality of Varicella-Zoster Virus Glycoprotein-Specific Antibodies Identified after Zostavax Vaccination in Humans. *J Virol* **92**, doi:10.1128/JVI.00269-18 (2018).

29 Michlmayr, D. *et al.* Comprehensive Immunoprofiling of Pediatric Zika Reveals Key Role for Monocytes in the Acute Phase and No Effect of Prior Dengue Virus Infection. *Cell Rep* **31**, 107569, doi:10.1016/j.celrep.2020.107569 (2020).

30 Katzelnick, L. C. *et al.* Antibody-dependent enhancement of severe dengue disease in humans. *Science* **358**, 929-932, doi:10.1126/science.aan6836 (2017).

31 Kazmin, D. *et al.* Systems analysis of protective immune responses to RTS,S malaria vaccination in humans. *Proc Natl Acad Sci U S A* **114**, 2425-2430, doi:10.1073/pnas.1621489114 (2017).

32 Mpina, M. *et al.* Controlled Human Malaria Infection Leads to Long-Lasting Changes in Innate and Innate-like Lymphocyte Populations. *J Immunol* **199**, 107-118, doi:10.4049/jimmunol.1601989 (2017).

33 Li, S. *et al.* Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nat Immunol* **15**, 195-204, doi:10.1038/ni.2789 (2014).

34 Team, H.-C. S. P. & Consortium, H.-I. Multicohort analysis reveals baseline transcriptional predictors of influenza vaccination responses. *Sci Immunol* **2**, doi:10.1126/sciimmunol.aal4656 (2017).

35 Azuaje, F. Computational models for predicting drug responses in cancer research. *Brief Bioinform* **18**, 820-829, doi:10.1093/bib/bbw065 (2017).

36 Jia, S., Li, J., Liu, Y. & Zhu, F. Precision immunization: a new trend in human vaccination. *Hum Vaccin Immunother* **16**, 513-522, doi:10.1080/21645515.2019.1670123 (2020).

37 Gao, A. *et al.* Predicting the Immunogenicity of T cell epitopes: From HIV to SARS-CoV-2. *bioRxiv*, doi:10.1101/2020.05.14.095885 (2020).

38 Chaussabel, D. Assessment of immune status using blood transcriptomics and potential implications for global health. *Semin Immunol* **27**, 58-66, doi:10.1016/j.smim.2015.03.002 (2015).

39 Bhattacharya, S. *et al.* ImmPort: disseminating data to the public for the future of immunology. *Immunol Res* **58**, 234-239, doi:10.1007/s12026-014-8516-1 (2014).

40 Kauffmann, A., Gentleman, R. & Huber, W. arrayQualityMetrics--a bioconductor package for quality assessment of microarray data. *Bioinformatics* **25**, 415-416, doi:10.1093/bioinformatics/btn647 (2009).

41 Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-264, doi:10.1093/biostatistics/4.2.249 (2003).

42 Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15**, R29, doi:10.1186/gb-2014-15-2-r29 (2014).

43 Bruford, E. A. *et al.* Guidelines for human gene nomenclature. *Nat Genet* **52**, 754-758, doi:10.1038/s41588-020-0669-3 (2020).

44 Boedigheimer, M. J. *et al.* Sources of variation in baseline gene expression levels from toxicogenomics study control animals across multiple laboratories. *BMC Genomics* **9**, 285, doi:10.1186/1471-2164-9-285 (2008).

45 Bulteau, R. & Francesconi, M. Real age prediction from the transcriptome with RAPToR. *bioRxiv*, 2021.2009.2007.459270, doi:10.1101/2021.09.07.459270 (2021).

46 Shah, N. *et al.* A crowdsourcing approach for reusing and meta-analyzing gene expression data. *Nat Biotechnol* **34**, 803-806, doi:10.1038/nbt.3603 (2016).

FIGURE LEGENDS:

Figure 1: HIPC Immune Signatures Data Resource pipeline and study demographics.

A. Systems vaccinology datasets from existing HIPC studies, as well as published systems vaccinology papers and databases, were submitted to the *ImmPort* database. *ImmuneSpace* captures these datasets to create a combined compendium dataset. Quality control assessments of these data include array quality checks for microarray studies, batch correction, imputations for missing age and sex/y-chromosome presence information, and normalization per study. The combined virtual study included transcriptional profiles and antibody response measurements from 1405 participants across 53 cohorts, profiling the response to 24 different vaccines.

B. Demographic data included biological sex, race, vaccine, and number of participants.

Figure 2: Flow chart diagram of the Immune Signatures Data Resource.

Figure 3: Quality control assessments of transcriptomics data.

A. Sample quality assessments of gene expression datasets using Array Quality metrics. Array quality metrics package was employed to assess quality of microarray datasets by checking the following criteria: a.) absolute mean difference between arrays to check the probe and median intensity across all arrays, b.) Kolmogorov-Smirnov statistics to check the signal intensity distribution of arrays, comparing each probe versus distribution of test statistics for all other probes, c.) Hoeffding's D-statistics for arrays. Arrays were excluded if they fail all three criteria above.

B-C: Principal component analysis (Top) and Principal Variation component Analysis (PVCA) of baseline expression data per study before (B) and after batch correction (C).

D. Biological sex imputation based on expression of Y-chromosome genes. We used 13 Y-chromosome-associated genes to cluster samples into 2 groups assuming biological male or female.

E-F. Age imputation based on transcriptomic profiles for studies without reported ages (SDY1260, SDY1264, SDY1293, SDY1294, SDY1364, SDY1370, SDY1373, SDY984) via the RAPToR R package[45]. Virtual studies were split into young (age < 50, **E**) and older (age >= 50, **F**) for two separate predictive models.

Figure 4: Immune Signatures Transcriptomics Overview for young and old datasets.

A. Number of samples available for each data type, including transcriptomics (TX), hemagglutination inhibition assay (HAI), neutralizing antibody assay (NAB), and ELISA assays (ELISA).

B. Bar plot depicting the number of samples at each time point. The colors within each bar indicate the breakdown for each unique combination of pathogen and vaccine type. Day -7 and day 0 correspond to times pre-vaccination.

C. Box plot depicting the participant's age distribution for each unique combination of pathogen and vaccine type D. Each area-proportional Euler diagram represents the total number of participants with corresponding data types.

Figure 5: Immune Response Dataset Overview

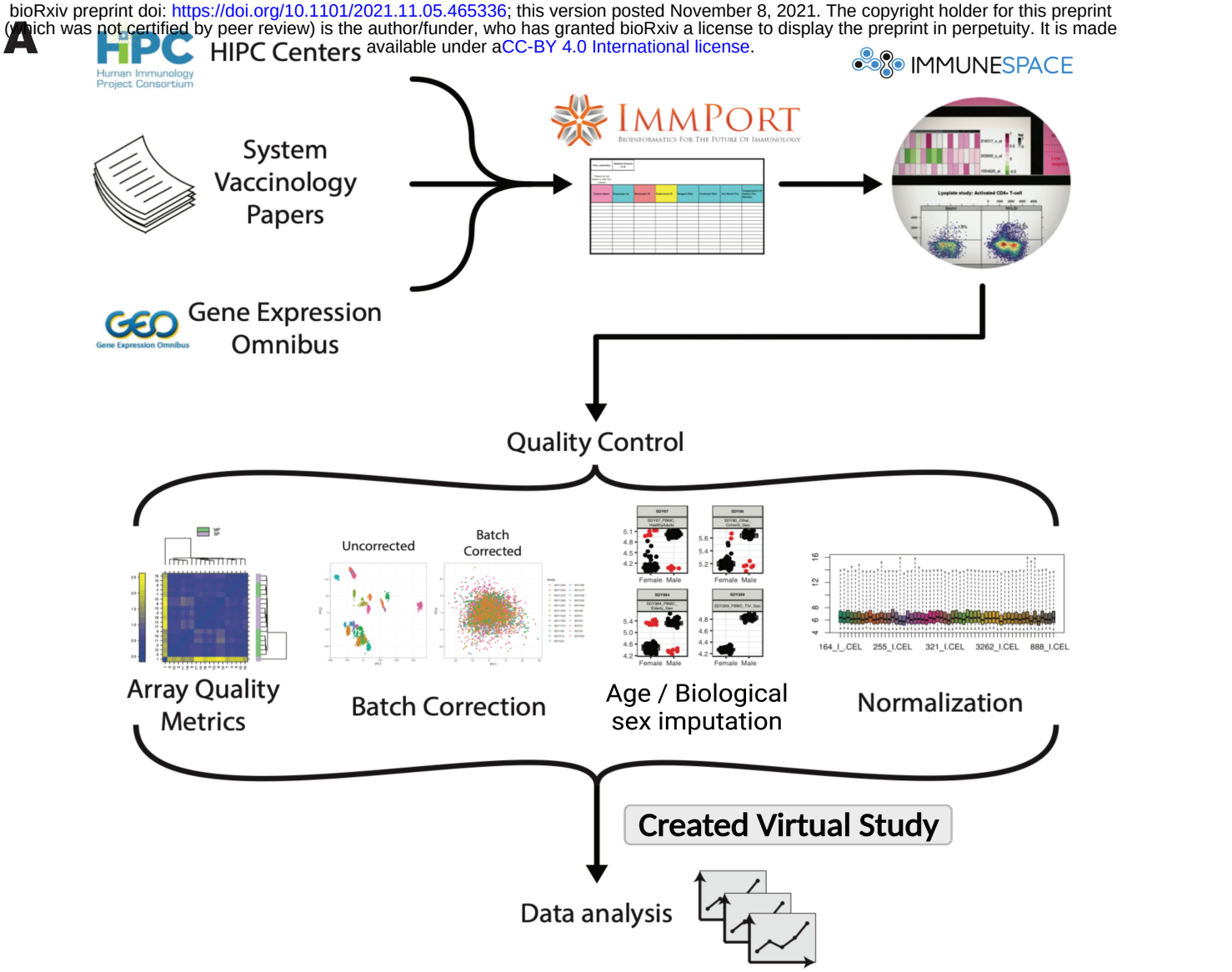
A. The longitudinal trajectory (summarized as a loess curve) of hemagglutinin inhibition assay (HAI) measurements (in $\log_2$ scale) by influenza vaccine type and year.

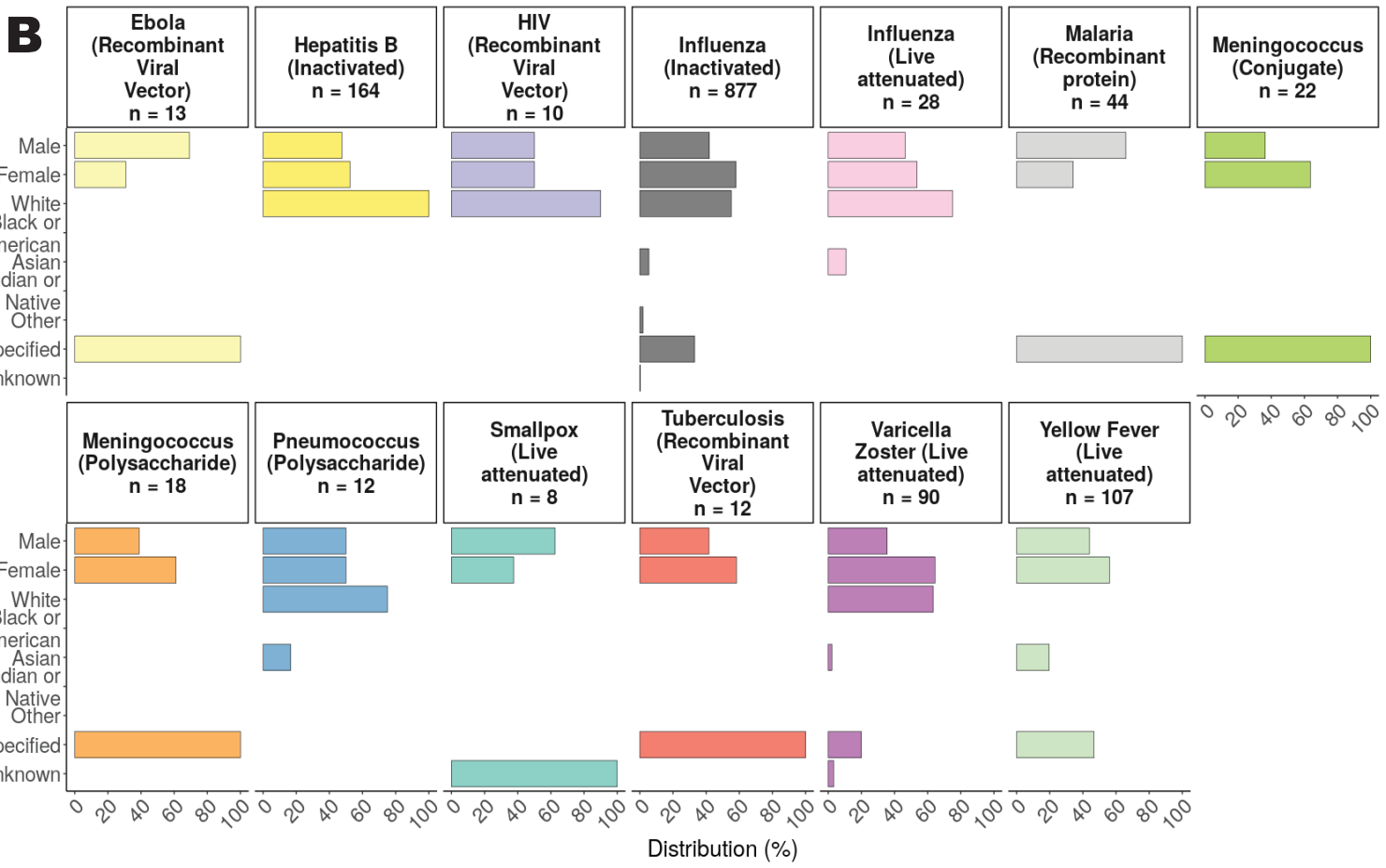
B. The longitudinal trajectory of neutralizing antibody (NAB) titers (in log2 scale) for influenza, meningococcus, pneumococcus, and yellow fever vaccines.

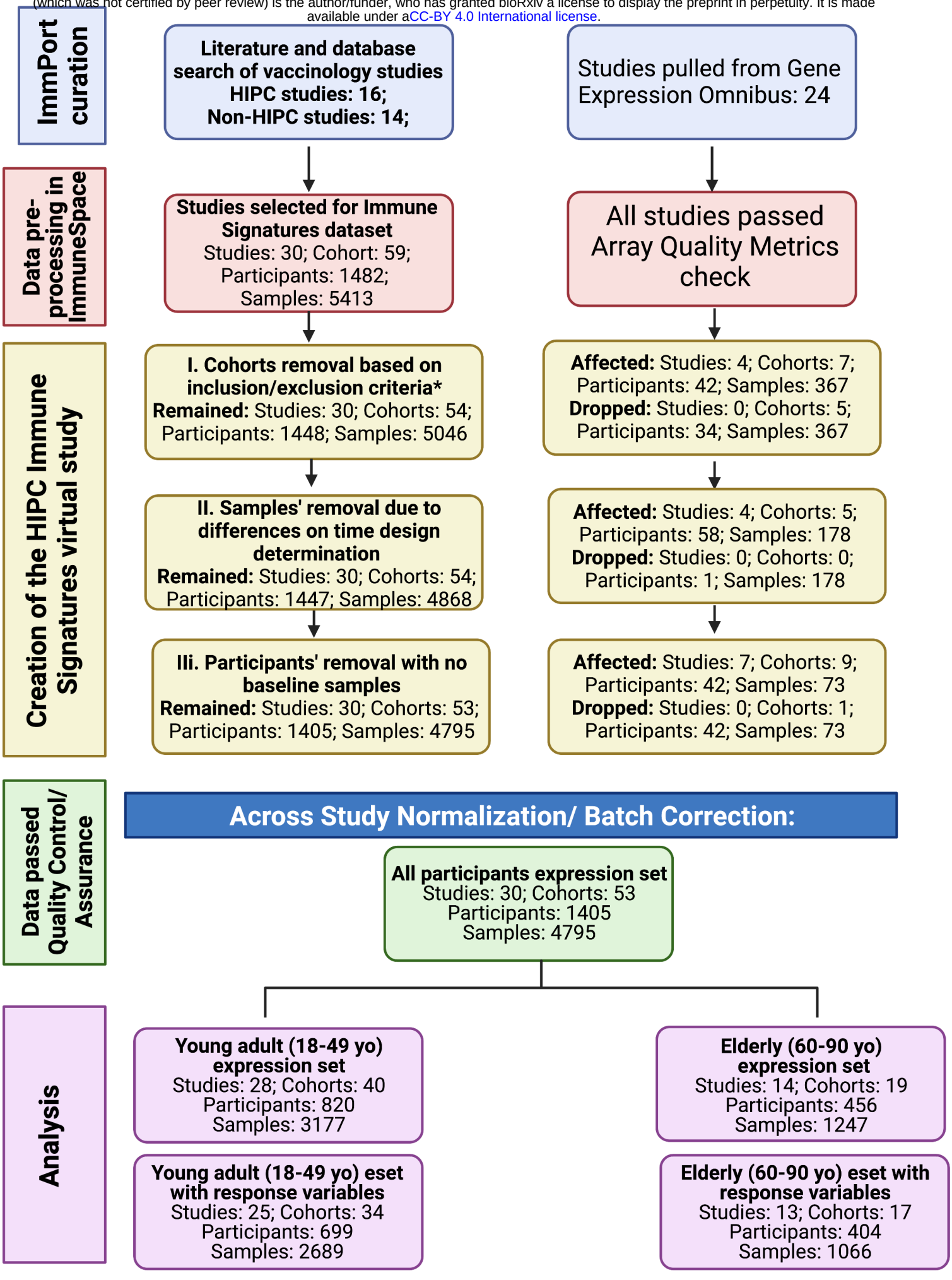
C. Neutralizing antibody titers were plotted for each unique combination of targeted pathogen and vaccine type to compare each participants' post-vaccination (day 28-30) values versus baseline (day 0). The violin plot shows the variation in magnitude for each unique combination of targeted pathogen and vaccine type.

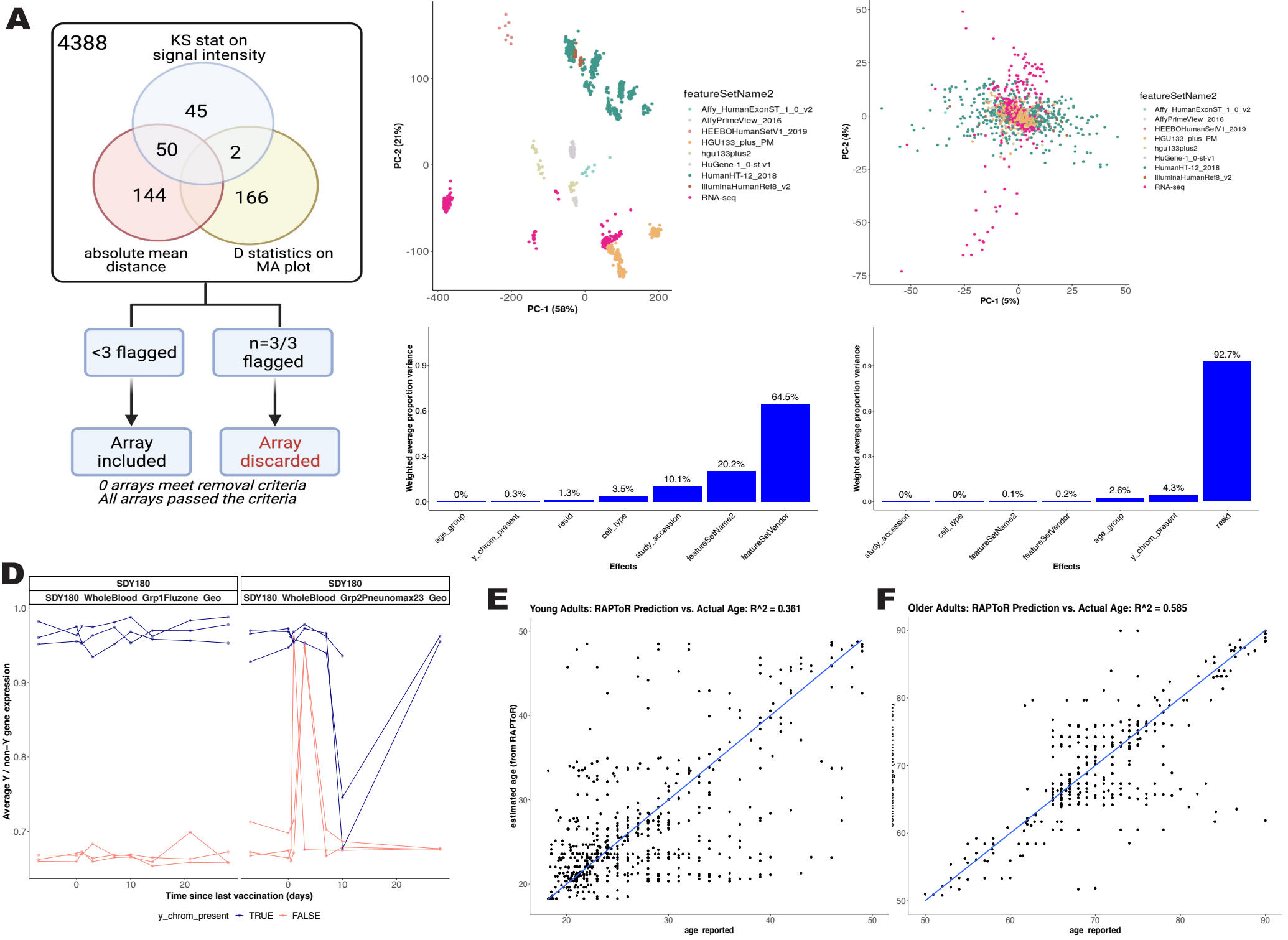
D. The correlation plot of influenza studies compares the maximum fold change (MFC) across strains for hemagglutinin inhibition assay (HAI) titers versus neutralizing antibody (NAB) titers. Size is proportional to the number of samples analyzed.

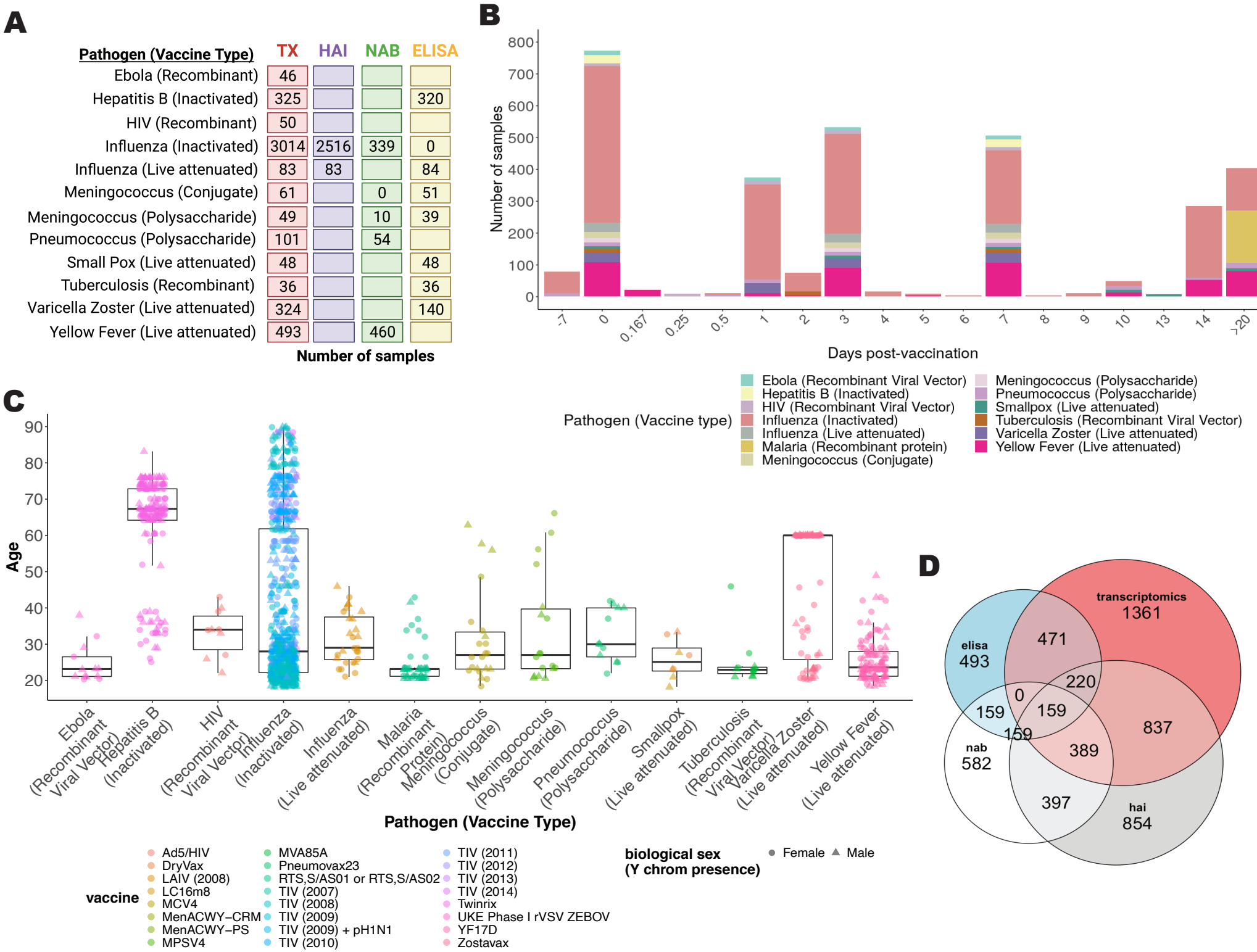
575

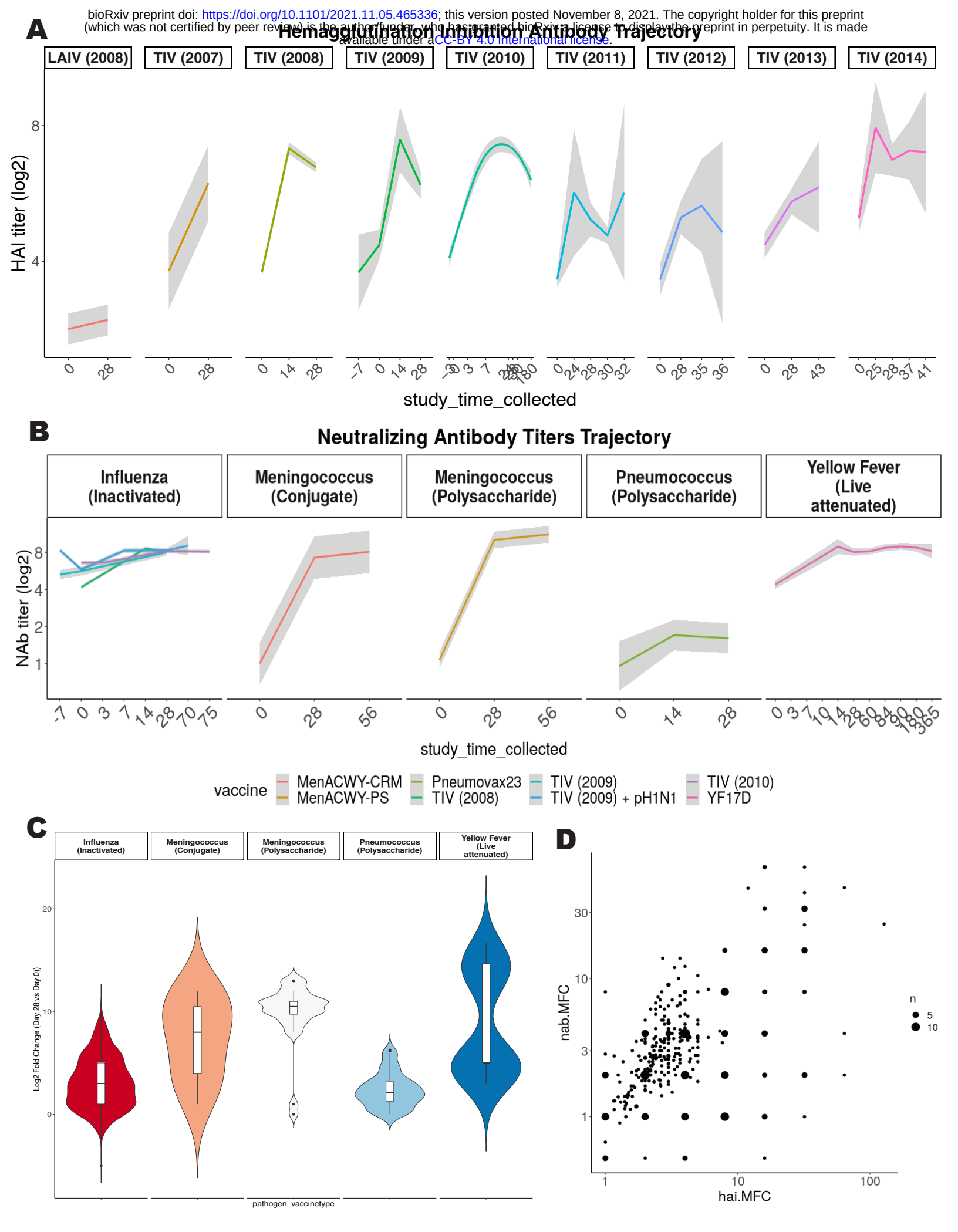







A
4388
KS stat on signal intensity
45
50
2
144
166
absolute mean distance
D statistics on MA plot
<3 flagged
n=3/3 flagged
Array included
Array discarded
0 arrays meet removal criteria
All arrays passed the criteria
B
Baseline Expression
PC-2 (21%)
PC-1 (58%)
featureSetName2
Affy_HumanExonST_1_0_v2
AffyPrimeView_2016
HEEBOHumanSetV1_2019
HGU133_plus_PM
hgu133plus2
HuGene-1_0-st-v1
HumanHT-12_2018
IlluminaHumanRef8_v2
RNA-seq
Weighted average proportion variance
0%
0.3%
1.3%
3.5%
10.1%
20.2%
64.5%
age_group
y_chrom_present
resid
cell_type
study_accession
featureSetName2
featureSetVendor
Effects
C
Baseline post-batch correction
PC-2 (4%)
PC-1 (5%)
Weighted average proportion variance
0%
0%
0.1%
0.2%
2.6%
4.3%
92.7%
study_accession
cell_type
featureSetName2
featureSetVendor
age_group
y_chrom_present
resid
Effects
D
SDY180
SDY180_WholeBlood_Grp1Fluzone_Geo
SDY180_WholeBlood_Grp2Pneunomax23_Geo
Average Y / non-Y gene expression
Time since last vaccination (days)
y_chrom_present TRUE FALSE
E
Young Adults: RAPToR Prediction vs. Actual Age: R^2 = 0.361
estimated age (from RAPToR)
age_reported
F
Older Adults: RAPToR Prediction vs. Actual Age: R^2 = 0.585
age_reported


A
Pathogen (Vaccine Type)
TX
HAI
NAB
ELISA
Ebola (Recombinant) 46
Hepatitis B (Inactivated) 325 320
HIV (Recombinant) 50
Influenza (Inactivated) 3014 2516 339 0
Influenza (Live attenuated) 83 83 84
Meningococcus (Conjugate) 61 0 51
Meningococcus (Polysaccharide) 49 10 39
Pneumococcus (Polysaccharide) 101 54
Small Pox (Live attenuated) 48 48
Tuberculosis (Recombinant) 36 36
Varicella Zoster (Live attenuated) 324 140
Yellow Fever (Live attenuated) 493 460
Number of samples
B
Number of samples
Days post-vaccination
C
Age
Pathogen (Vaccine Type)
Pathogen (Vaccine type)
Ebola (Recombinant Viral Vector)
Hepatitis B (Inactivated)
HIV (Recombinant Viral Vector)
Influenza (Inactivated)
Influenza (Live attenuated)
Malaria (Recombinant protein)
Meningococcus (Conjugate)
Meningococcus (Polysaccharide)
Pneumococcus (Polysaccharide)
Smallpox (Live attenuated)
Tuberculosis (Recombinant Viral Vector)
Varicella Zoster (Live attenuated)
Yellow Fever (Live attenuated)
vaccine
Ad5/HIV
DryVax
LAIV (2008)
LC16m8
MCV4
MenACWY−CRM
MenACWY−PS
MPSV4
MVA85A
Pneumovax23
RTS,S/AS01 or RTS,S/AS02
TIV (2007)
TIV (2008)
TIV (2009)
TIV (2009) + pH1N1
TIV (2010)
TIV (2011)
TIV (2012)
TIV (2013)
TIV (2014)
Twinrix
UKE Phase I rVSV ZEBOV
YF17D
Zostavax
biological sex (Y chrom presence)
Female
Male
D
transcriptomics 1361
elisa 493
nab 582
hai 854
471
220
0
159
159
159
837
389
397



1 **TABLES:**

2 **Table 1: Overview of Immune Signatures Data Resource Study Participants Metadata**

Study Accession	Pathogen (Vaccine Type)	Number of Participants	Number of Samples	Vaccine	Adjuvant	Race	Ethnicity	Cohort	Matrix	Pubmed ID
SDY1373	Ebola (Recombinant Viral Vector)	13	46	UKE Phase I rVSV ZEBOV	VSV	Not Specified	Not Specified	dose 20x10^6 ofu,dose 3x10^6 pfu	SDY1373_WholeBlood_highDose_Geo,SDY1373_WholeBlood_lowDose_Geo	
SDY1328	Hepatitis B (Inactivated)	164	325	Twinrix	None	White	Not Hispanic or Latino	healthy adults	SDY1328_WholeBlood_HealthyAldults_Geo	26742691
SDY1291	HIV (Recombinant Viral Vector)	10	50	Ad5/HIV	AdV	White, Black, or African American	Not Hispanic or Latino	healthy HIV-1-uninfected adults	SDY1291_PBMC_HealthyHIVUninfected_Geo	23151505
SDY1119	Influenza (Inactivated)	72	177	TIV (2011)	None	Not Specified	Not Specified	young T2D, young healthy,old healthy old T2D	SDY1119_PBMC_youngT2D_Geo,SDY1119_PBMC_youngHealthy_Geo,SDY1119_PBMC_oldHealthy_Geo,SD	26682988

									Y1119_PBMC_oldT2D_Geo	
SDY1276	Influenza (Inactivated)	218	828	TIV (2008)	None	Not Specified	Not Specified	Validation Cohort; Females 2008-2009 trivalent influenza vaccine ,Discovery Cohort; Males 2008?2009 trivalent influenza vaccine	SDY1276_WholeBlood_Validation_Geo,SDY1276_WholeBlood_Discovery_Geo	21357945
SDY180	Influenza (Inactivated)	12	102	TIV (2009)	None	Asian ,White,Black or African American	Not Hispanic or Latino	Study group 2 2009-2010 Fluzone,Study group 1 2009-2010 Fluzone	SDY180_WholeBlood_Grp2Fluzone_Geo,SDY180_WholeBlood_Grp1Fluzone_Geo	23601689
SDY212	Influenza (Inactivated)	90	90	TIV (2008)	None	Oth er,White,Asian,American I,ndian or Alaska Native	Not Hispanic or L atino,Hispanic or Latino	Cohort_1,Cohort_2	SDY212_WholeBlood_Young_Geo,SDY212_PBMC_Young_geo,SDY212_WholeBlood_Older_G	23591775

									eo,SDY212_PBMC_Older_Geo	
SDY224	Influenza (Inactivated)	5	55	TIV (2010)	None	White,Black or African American, American Indian or Alaska Native	Not Hispanic or Latino,Hispanic or Latino	TIV 2010	SDY224_PBMC_TIV 2010_*ImmPort*	23900141
SDY269	Influenza (Inactivated)	28	80	TIV (2008)	None	White,Asian,Black or African American	Not Hispanic or Latino,Hispanic or Latino	TIV Group 2008	SDY269_PBMC_TIV_Geo	21743478
SDY270	Influenza (Inactivated)	28	83	TIV (2009)	None	White,Black or African American, Asian	Not Hispanic or Latino,Hispanic or Latino	TIV Group 2009	SDY270_PBMC_TIV Group_Geo	21743478
SDY400	Influenza (Inactivated)	30	120	TIV (2012)	None	White,Asian,Black or African American, Other	Not Hispanic or Latino,Hispanic or Latino	Young adults 21-30 years old,Older adults >= 65 years old	SDY400_PBMC_Young_Geo,SDY400_PBMC_Older_Geo	32060136
SDY404	Influenza (Inactivated)	39	156	TIV (2011)	None	White,Unknown,Other,Asian,Black or African	Not Hispanic or Latino,Hispanic or Latino	Young adults 21-30 years old,Older adults >=	SDY404_PBMC_Young_Geo,SDY404_PBMC_Older_G	25596819

						American		65 years old	eo	
SDY520	Influenza (Inactivated)	24	94	TIV (2013)	None	White,Asian,Black or African American	Not Hispanic or Latino,Hispanic or Latino	Young adults 21-30 years old,Older adults >= 65 years old	SDY520_WholeBlood_Young_geo,SDY520_WholeBlood_Older_Geo	32060136
SDY56	Influenza (Inactivated)	63	288	TIV (2010)	None	White,Asian,Black or African American	Not Hispanic or Latino,Hispanic or Latino	Healthy adults 25-40 years old receiving TIV flu vaccine,Healthy adults >65 years old receiving TIV flu vaccine	SDY56_PBMC_Young,SDY56_PBMC_Older	26682988
SDY61	Influenza (Inactivated)	9	27	TIV (2007)	None	White	Not Hispanic or Latino,Hispanic or Latino	TIV Group 2007	SDY61_PBMC_TIVGrp	21743478
SDY63	Influenza (Inactivated)	19	72	TIV (2010)	None	White,Asian,Other,Black or African American	Not Hispanic or Latino	Young adults 21-30 years old,Older adults >= 65 years old	SDY63_PBMC_Young_Geo,SDY63_PBMC_Older_Geo	25596819

SDY640	Influenza (Inactivated)	20	79	TIV (2014)	None	White,Asian,Unknown	Not Hispanic or Latino,Hispanic or Latino	Young adults 21-30 years old,Older adults >= 65 years old	SDY640_WholeBlood_Young_Geo,SDY640_WholeBlood_Older_Geo	32060136
SDY80	Influenza (Inactivated)	61	286	TIV (2009) + pH1N1	None	White,Asian,Other,Black or African American	Other,Hispanic or Latino	Cohort2	SDY80_PBMC_Cohort2_geo	24725414
SDY269	Influenza (Live attenuated)	28	83	LAIV (2008)	LAIV	White,Black or African American, Asian	Not Hispanic or Latino,Hispanic or Latino	LAIV group 2008	SDY269_PBMC_LAIV_Geo	21743478
SDY1293	Malaria (Recombinant protein)	44	165	RTS,S/AS01 or RTS,S/AS02	AS01/AS02	Not Specified	Not Specified	adjuvanted RTS,S malaria vaccine cohort	SDY1293_PBMC_Vaccinated_geo	
SDY1260	Meningococcus (Conjugate)	17	51	MCV4	None	Not Specified	Not Specified	MCV4	SDY1260_PBMC_MCV4_Geo	24336226
SDY1325	Meningococcus (Conjugate)	5	10	MenACWY-CRM	None	Not Specified	Not Specified	Intramuscular MenACWY-CRM	SDY1325_WholeBlood_IntramuscularCRM_Geo	28137280
SDY1260	Meningococcus (Polysaccharide)	13	39	MPSV4	None	Not Specified	Not Specified	MPSV4	SDY1260_PBMC_MPSV4_Geo	24336226

SDY1325	Meningococcus (Polysaccharide)	5	10	MenACWY-PS	None	Not Specified	Not Specified	Intramuscular MenACWY-PS	SDY1325_WholeBlood_IntramuscularPS_Geo	28137280
SDY180	Pneumococcus (Polysaccharide)	12	101	Pneumovax23	None	White,Black or African American, Asian	Not Hispanic or Latino,Hispanic or Latino	Study group 2 Pneunomax 23,Study group 1 Pneunomax 23	SDY180_WholeBlood_Grp2Pneunomax23_Geo,SDY180_WholeBlood_Grp1Pneunomax23_Geo	23601689
SDY1370	Smallpox (Live attenuated)	4	24	DryVax	Vaccinia	Unknown	Not Specified	DryVax	SDY1370_PBMC_dryvax_geo	21921208
SDY1370	Smallpox (Live attenuated)	4	24	LC16m8	Vaccinia	Unknown	Not Specified	LC16m8	SDY1370_PBMC_lc16m8_geo	21921208
SDY1364	Tuberculosis (Recombinant Viral Vector)	12	36	MVA85A	Vaccinia	Not Specified	Not Specified	MVA85A intramuscular	SDY1364_PBMC_IntraMuscular_Geo	23844129
SDY984	Varicella Zoster (Live attenuated)	72	288	Zostavax	VZV	White,Black or African American, Unknown, Asian	Not Hispanic or Latino,Hispanic or Latino	young,elderly	SDY984_PBMC_Young_Geo,SDY984_PBMC_Elderly_Geo	28502771
SDY1264	Yellow Fever (Live attenuated)	25	87	YF17D	YF17D	Not Specified	Not Specified	Trial2,Trial1	SDY1264_PBMC_Trial2_Geo,S	19029902

									DY1264_PBMC_Trial1_Geo	
SDY1289	Yellow Fever (Live attenuated)	25	117	YF17D	YF17D	Not Specified	Not Specified	in vivo vaccination study Montreal adult cohort,in vivo vaccination study Lausanne adult cohort	SDY1289_WholeBlood_MontrealCohort_Geo,SDY1289_WholeBlood_LausanneCohort_Geo	19047440
SDY1294	Yellow Fever (Live attenuated)	21	109	YF17D	YF17D	Asian	Not Hispanic or Latino	Chinese cohort	SDY1294_PBMC_ChineseCohort_Geo	28687661
SDY1529	Yellow Fever (Live attenuated)	36	180	YF17D	YF17D	Black or African American	Not Hispanic or Latino	healthy adults	SDY1529_WholeBlood_HealthyAdults_PreVax_Geo,SDY1529_WholeBlood_HealthyAdults_PostVax_Geo	19047440

7

8 **Table 2: Overview of Transcriptomics Datasets Included in the Resource**

Study Accession	Pathogen (Vaccine type)	Sample type	featureSetName	featureSetName2	featureSetVendor	Time post last vaccination	GEO Accession
SDY1373	Ebola (Recombinant Viral Vector)	Whole blood	SDY1373_customAnno	RNA-seq	NA	0,1,3,7	GSE97590
SDY1328	Hepatitis B (Inactivated)	Whole blood	Affy_HumanRSTAcustom	RNA-seq	Affymetrix	0,7	GSE65834
SDY1291	HIV (Recombinant Viral Vector)	PBMC	Affy_HumanExonST_1_0_v2	Affy_HumanExonST_1_0_v2	Affymetrix	0,0.25,1,3,7	GSE22768
SDY1119	Influenza (Inactivated)	PBMC	HGU133_plus_PM	HGU133_plus_PM	Affymetrix	0,3,7	GSE74817
SDY1276	Influenza (Inactivated)	Whole blood	HumanHT-12_v3_2018	HumanHT-12_2018	Illumina	0,1,3,14	GSE48024/GSE48018
SDY180	Influenza (Inactivated)	Whole blood	HumanHT-12_v3_2018	HumanHT-12_2018	Illumina	-7,0,0.5,1,3,7,10,14,21,28	GSE48762
SDY212	Influenza (Inactivated)	Whole blood	HumanHT-12_v3_2018	HumanHT-12_2018	Illumina	0	GSE41080
SDY224	Influenza (Inactivated)	PBMC	SDY224_CustomAnno	RNA-seq	NA	0,1,2,3,4,5,6,7,8,9,10	GSE45735
SDY269	Influenza (Inactivated)	PBMC	HGU133_plus_PM	HGU133_plus_PM	Affymetrix	0,3,7	GSE29615/GSE29617/GSE29614
SDY270	Influenza (Inactivated)	PBMC	HGU133_plus_PM	HGU133_plus_PM	Affymetrix	0,3,7	GSE29617/GSE29614
SDY400	Influenza (Inactivated)	PBMC	HumanHT-12_v4_2018	HumanHT-12_2018	Illumina	0,2,4,7,28	GSE59743/GSE95584
SDY404	Influenza (Inactivated)	PBMC	HumanHT-12_v4_2018	HumanHT-12_2018	Illumina	0,2,4,7,28	GSE59654
SDY520	Influenza (Inactivated)	Whole blood	HumanHT-12_v4_2018	HumanHT-12_2018	Illumina	0,2,7,28	GSE101709

SDY56	Influenza (Inactivated)	PBMC	HGU133_plus_PM	HGU133_plus_PM	Affymetrix	0,1,3,7,14	GSE74817
SDY61	Influenza (Inactivated)	PBMC	hgu133plus2	hgu133plus2	Affymetrix	0,3,7	GSE29617/GSE29614
SDY63	Influenza (Inactivated)	PBMC	HumanHT-12_v4_2018	HumanHT-12_2018	Illumina	0,4,7,28	GSE59635
SDY640	Influenza (Inactivated)	Whole blood	HumanHT-12_v4_2018	HumanHT-12_2018	Illumina	0,2,7,28	GSE101710
SDY80	Influenza (Inactivated)	PBMC	HuGene-1_0-st-v1	HuGene-1_0-st-v1	Affymetrix	-7,0,1,7,70	GSE47353
SDY269	Influenza (Live attenuated)	PBMC	HGU133_plus_PM	HGU133_plus_PM	Affymetrix	0,3,7	GSE29615/GSE29617/GSE29614
SDY1293	Malaria (Recombinant protein)	PBMC	hgu133plus2	hgu133plus2	Affymetrix	0,1,3,14	GSE18323
SDY1260	Meningococcus (Conjugate)	PBMC	HGU133_plus_PM	HGU133_plus_PM	Affymetrix	0,3,7	GSE52245
SDY1325	Meningococcus (Conjugate)	Whole blood	HumanHT-12_v4_2018	HumanHT-12_2018	Illumina	0,7	GSE92884
SDY1260	Meningococcus (Polysaccharide)	PBMC	HGU133_plus_PM	HGU133_plus_PM	Affymetrix	0,3,7	GSE52245
SDY1325	Meningococcus (Polysaccharide)	Whole blood	HumanHT-12_v4_2018	HumanHT-12_2018	Illumina	0,7	GSE92884
SDY180	Pneumococcus (Polysaccharide)	Whole blood	HumanHT-12_v3_2018	HumanHT-12_2018	Illumina	-7,0,0.5,1,3,7,10,14,21,28	GSE48762
SDY1370	Smallpox (Live attenuated)	PBMC	HEEBOHumanSetV1_2019	HEEBOHumanSetV1_2019	Stanford Functional Genomics Facility	0,3,7,10,13,21	GSE22121
SDY1370	Smallpox (Live attenuated)	PBMC	HEEBOHumanSetV1_2019	HEEBOHumanSetV1_2019	Stanford Functional Genomics	0,3,7,10,13,21	GSE22121

					Facility		
SDY1364	Tuberculosis (Recombinant Viral Vector)	PBMC	HumanHT-12_v4_2018	HumanHT-12_2018	Illumina	0,2,7	GSE40719
SDY984	Varicella Zoster (Live attenuated)	PBMC	HGU133_plus_PM	HGU133_plus_PM	Affymetrix	0,1,3,7	GSE79396
SDY1264	Yellow Fever (Live attenuated)	PBMC	hgu133plus2	hgu133plus2	Affymetrix	0,1,3,7,21	GSE13485
SDY1289	Yellow Fever (Live attenuated)	Whole blood	IlluminaHumanRef8_v2	IlluminaHumanRef8_v2	Illumina	0,3,7,10,14,28,60	GSE13699
SDY1294	Yellow Fever (Live attenuated)	PBMC	AffyPrimeView_2016	AffyPrimeView_2016	Affymetrix	0,0.166666666666667,1,2,3,5,7,14,28	GSE82152
SDY1529	Yellow Fever (Live attenuated)	Whole blood	HumanHT-12_v4_2018	HumanHT-12_2018	Illumina	0,3,7,14,84	GSE125921/GSE136163

Table 3: Studies with corresponding Immune Response Data

Study Accession	Pathogen Vaccine type	Number of Participants	Number of Samples	Selected Immune Response Assay
SDY1328	Hepatitis B (Inactivated)	160	320	ELISA
SDY1119	Influenza (Inactivated)	72	177	HAI
SDY1276	Influenza (Inactivated)	214	816	HAI, NAb
SDY180	Influenza (Inactivated)	12	102	HAI, NAb
SDY212	Influenza (Inactivated)	88	88	HAI
SDY224	Influenza (Inactivated)	5	55	HAI
SDY269	Influenza (Inactivated)	28	80	HAI
SDY270	Influenza (Inactivated)	28	83	HAI
SDY400	Influenza (Inactivated)	30	120	HAI
SDY404	Influenza (Inactivated)	39	156	HAI
SDY520	Influenza (Inactivated)	24	94	HAI
SDY56	Influenza (Inactivated)	30	148	HAI
SDY61	Influenza (Inactivated)	9	27	HAI
SDY63	Influenza (Inactivated)	19	72	HAI
SDY640	Influenza (Inactivated)	20	79	HAI
SDY67	Influenza (Inactivated)	159	477	HAI
SDY80	Influenza (Inactivated)	60	281	NAb
SDY269	Influenza (Live attenuated)	28	83	HAI
SDY1260	Meningococcus (Conjugate)	17	51	ELISA
SDY1325	Meningococcus (Conjugate)	4	8	NAb
SDY1260	Meningococcus (Polysaccharide)	13	39	ELISA
SDY1325	Meningococcus (Polysaccharide)	5	10	NAb
SDY180	Pneumococcus (Polysaccharide)	6	54	NAb
SDY1370	Smallpox (Live attenuated)	4	24	ELISA
SDY1370	Smallpox (Live attenuated)	4	24	ELISA
SDY1364	Tuberculosis (Recombinant Viral Vector)	12	36	ELISA
SDY984	Varicella Zoster (Live attenuated)	35	140	ELISA
SDY1264	Yellow Fever (Live attenuated)	25	87	NAb
SDY1289	Yellow Fever (Live attenuated)	14	84	NAb
SDY1294	Yellow Fever (Live attenuated)	21	109	NAb
SDY1529	Yellow Fever (Live attenuated)	36	180	NAb

13 **Table 4: List of data files for the Immune Signatures Data Resource**

File name	Description
all_noNorm_eset.rds	Gene expression matrix of all participants, log2-normalized expression
all_noNorm_withResponse_eset.rds	Gene expression matrix of all participants with matched immune response data, log2-normalized expression
all_norm_eset.rds	Gene expression matrix of all participants that are cross-study normalized and batch corrected
all_norm_withResponse_eset.rds	Gene expression matrix of all participants with matched simmune response dataset, cross-study normalized and batch corrected
young_noNorm_eset.rds	Gene expression matrix of participants aged 18-50, log2-normalized
young_noNorm_withResponse_eset.rds	Gene expression matrix of participants aged 18-50 with matched immune response data, log2-normalized
young_norm_eset.rds	Gene expression matrix of participants aged 18-50, cross-study normalized and batch corrected
young_norm_withResponse_eset.rds	Gene expression matrix of participants aged 18-50 with matched immune response data, cross-study normalized and batch corrected
old_noNorm_eset.rds	Gene expression matrix of participants aged 60-90, log2-normalized
old_noNorm_withResponse_eset.rds	Gene expression matrix of participants aged 60-90 with matched immune response data, log2-normalized expression
old_norm_batchCorrectedFromYoung_eset.rds	Gene expression matrix of participants aged 60-90, cross-study normalized and batch corrected using age correction coefficients from young
old_norm_batchCorrectedFromYoung_withResponse_eset.rds	Gene expression matrix of participants aged 60-90 with matched immune response data, cross-study normalized and batch corrected using age correction coefficients from young
extendedOld_noNorm_eset.rds	Gene expression matrix of participants aged 50-90, log2-normalized expression
extendedOld_noNorm_withResponse_eset.rds	Gene expression matrix of participants aged 50-90 with matched immune response data, log2-normalized counts

extendedOld_norm_batchCorrectedFromYoung_eset.rds	Gene expression matrix of participants aged 50-90, log2-normalized expression
extendedOld_norm_batchCorrectedFromYoung_withResponse_eset.rds	Gene expression matrix of participants aged 50-90 with immune response data, cross-study normalized, and batch corrected using correction coefficients from young