The FluPRINT dataset: A multidimensional analysis of the influenza vaccine imprint on the immune system

Authors

Adriana Tomic[1,2*], Ivan Tomic[3], Cornelia L. Dekker[4], Holden T. Maecker[5] and Mark M. Davis[1,6,7]

Affiliations

[1]Institute of Immunity, Transplantation and Infection, Stanford University School of Medicine, Stanford, CA 94304, USA.

[2]Oxford Vaccine Group, Department of Pediatrics, University of Oxford, Oxford OX3 9DU, UK.

[3]Independent Researcher, Electronic address: info@ivantomic.com.

[4]Department of Pediatrics, Stanford University School of Medicine, Stanford, CA 94304, USA.

[5]Human Immune Monitoring Center, Stanford University, Stanford, CA 94304, USA.

[6]Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA 94304, USA.

[7]Howard Hughes Medical Institute, Stanford University, Stanford, CA 94304, USA.

*Corresponding author: Adriana Tomic (atomic@stanford.edu)

Abstract (168 words)

Recent advances in machine learning have allowed identification of molecular and cellular factors that underly successful antibody responses to influenza vaccines. Results of these studies have revealed the high level of complexity necessary to establish influenza immunity, and many different cellular and molecular components involved. However, identified correlates of protection fail to account for the majority of vaccinated cases across ages, cohorts, and influenza seasons. Major challenges arise from small sample sizes and from analysis of only one aspect of the biology such by using transcriptome data. The objective of the current study is to create a unified database, entitled FluPRINT, to enable a large-scale study exploring novel cellular and molecular underpinnings of successful immunity to influenza vaccines. Over 3,000 parameters were considered, including serological responses to influenza strains, serum cytokines, cell subset phenotypes, and cytokine stimulations. FluPRINT, thus facilitates application of machine learning algorithms for data mining. The data are publicly available and represent a resource to uncover new markers and mechanisms that drive successful influenza vaccination.

Background and Summary (394 words)

Influenza virus has a devastating societal impact, causing up to 650,000 deaths every year worldwide[1]. Vaccination can prevent influenza-like illnesses, and thus lower the risk of the virus outbreak. However, currently available vaccines do not always provide protection, even among otherwise-healthy people, leading to serious pandemics. Application of computational biology and machine learning to clinical datasets holds promise for identifying immune cell populations and genes that mediate stronger antibody responses to influenza vaccines[2-9]. Lack of pre-existing influenza immunity, especially T cells, has been identified as one of the major predispositions for failure to generate effective response to vaccination[10-12]. However, exact phenotypes of $CD4^+$ and $CD8^+$ T cells, which are important for protective influenza immunity in general and to vaccination with live attenuated influenza vaccine (LAIV) in specific, remain elusive. Identified correlates of protection, moreover, are not consistent between cohorts and study years. Some of the identified challenges leading to such discrepancy are small sample sizes and analysis of only one aspect of the biology, such as molecular correlates of protection by using transcriptome data[13]. Additionally, comparison of the results of different predictive models is hampered by the lack of a consensus regarding what defines the outcome of vaccination, i.e. high vs. low responders. For these reasons, it is necessary to generate a unified dataset that includes multiple measurements across age, gender, and racially diverse populations.

Specifically, it is of the utmost importance to include single-cell analysis at the protein level, such as mass cytometry combined with multiple high-dimensional biological measurements, since these have power to reveal heterogeneity of the immune system[14-18].

To accomplish that goal, we generated FluPRINT, a dataset consisting of 14 data types in standardized tables on blood and serum samples taken from 747 individuals undergoing influenza vaccination (**Fig. 1**). The FluPRINT dataset contains information on more than 3,000 parameters measured using mass cytometry (CyTOF), flow cytometry, phosphorylation-specific cytometry (phospho-flow), multiplex ELISA (Luminex assays), clinical lab tests (hormones and complete blood count), serological profiling with hemagglutination inhibition assay (HAI), and virological tests. The dataset represents fully integrated and normalized immunology measurements from eight clinical studies conducted between 2007 to 2015 at the Human Immune Monitoring Center (HIMC) of Stanford University. The MySQL database containing this immense dataset is publicly available online (www.fluprint.com). The dataset represents a unique source in terms of value and scale, which will broaden our understanding of influenza immunity.

Methods

Clinical studies. All studies were approved by the Stanford Institutional Review Board and performed in accordance with guidelines on human cell research. Peripheral blood samples were obtained at the Clinical and Translational Research Unit at Stanford University after written informed consent/assent was obtained from participants. Samples were processed and cryopreserved by the Stanford HIMC BioBank according to the standard operating protocols[19].

Data collection. Data involving individuals enrolled in influenza vaccine studies at the Stanford-LPCH Vaccine Program was accessed from the Stanford Data Miner (SDM) which holds data processed by HIMC from 2007 up to date[20]. The FluPRINT cohort was assembled by filtering the SDM for assays available in studies involving influenza vaccination. This resulted in a dataset containing data from 820 healthy donors enrolled in influenza vaccine studies at the Stanford-LPCH Vaccine Program from 2007 to 2015 in the following studies: SLVP015 (NCT01827462 and NIAID ImmPort accession number SDY212) part of the data published in[4], SLVP017 (NCT02133781, NCT03020498, NCT03020537), SLVP018 (NCT01987349, NCT03022396, NCT03022422, NCT03022435, NCT3023176) data partially published in[21], SLVP021 (NCT02141581), SLVP024 (NCT03023683), SLVP028 (NCT03088904), SLVP029 (NCT03028974) and SLVP030 (NCT03453801). A total of

177 CSV files containing raw data from various assays were downloaded. The link to the 177 CSV files is provided as **Data Citation 1**. **Table 1** provides a summary of the demographic characteristics of the FluPRINT study population. The population spans a wide age range, from a 1-year-old to a 90-year-old, with a median age of 27 years. Among 747 individuals with available experimental data, 452 were females and 295 males. The majority (497) of the individuals were of European ancestry. The complete demographic information is provided as the **Data Citation 2**. Additionally, individuals were stratified into high and low responders, depending on their HAI antibody titers measured before and after vaccination, as described below. **Figure 2** shows demographic information for the FluPRINT study population, including gender, ethnicity, cytomegalovirus (CMV) status, and age stratified by the outcome to vaccination. Out of 364 individuals with measured HAI responses, 111 were identified as high responders and 253 as low responders. Overall, no major differences in the gender, ethnicity distribution, or CMV status (**Fig. 2a**) or age (**Fig. 2b**) were observed between high and low responders.

Assays performed. All data used were analysed and processed at the HIMC, as previously described[22]. Protocols for all assays performed such as immunophenotyping and determination of signalling responses to stimulation using flow or mass cytometry, HAI titer determination and analysis of cytokines/chemokines in samples using Luminex assay are available online[19]. Immunophenotyping using mass cytometry was published in Leipold and Maecker[23]. Previous studies describe: phospho-flow assay using flow cytometry (for studies SLVP015, SLVP018 and SLVP021 from 2007 to 2012)[4,21], mass cytometry (for studies SLVP015, SLVP018 and SLVP021 in 2013)[24] and Luminex assay[4,21]. The HAI assay was performed on sera from day 0 and day 28 using a well-established method[25] as previously described[3,4]. The distribution of assays performed across clinical studies and years is illustrated in **Table 2**. Overall, SLVP015 was the longest study, running from 2007 to 2014, spanning 135 unique individuals, while the majority of samples (256) came from the SLVP018 study (**Fig. 1**).

Automated importer and data harmonization. After collecting the data, a custom PHP script was generated to parse each of the 177 CSV files and to import data into the MySQL database. The source code for the script is available online at https://github.com/LogIN-/fluprint. The script optimizes the data harmonization process essential for combining data from different studies. Control and nonsense data were not imported, such as "CXCR3-FMO CD8+ T cells", "nonNK-nonB-nonT-nonmonocyte-nonbasophils", "viable", etc. To standardize data, the original CSV entries were cleaned into the MySQL database readable format (e.g. quotes and

parenthesis replaced with underscores, “+” with text “positive”, etc.). Additionally, classifications for ethnicity, vaccine names and age were resolved into standard forms. For example, "Fluzone single-dose syringe" and "Fluzone single-dose syringe 2009-2010" were mapped to “Fluzone”. During data merging, we replaced text strings with binary values. For example, for the variable of gender, female and male were replaced with zero and one. To be able to distinguish between visits in consecutive years, an internal visit identification was calculated. For the original internal visit data, each visit in one year was labelled as V1 for day zero and V2 for day seven. However, if the same individual came in the consecutive year, day zero visit would again be labelled V1, and day seven as visit V2, causing repetition of values. To avoid such repetitions in the database, we also generated a unique visit ID. Therefore, for the above example, first visit in the first year would be labelled V1 for day zero and V2 for day seven, but for the next year visits would be labelled as V3 for day zero and V4 for day seven. To distinguish between Luminex assays, the prefix L50 was given to each analyte analysed with the 51/52-plex Luminex kit. Finally, we imputed new values and calculated the vaccine outcome parameter using HAI antibody titers. High responders were determined as individuals that have HAI antibody titer for all vaccine strains above 40 and geometric mean (GeoMean) HAI fold change > 4. The fold change is calculated as: GeoMean HAI antibody titer for all vaccine strains on day 28 / GeoMean HAI antibody titer for all vaccine strains on day zero. Vaccine outcome was expressed as a binary value: high responders were given value of one and low responders the value zero.

Generating tables. To build FluPRINT database, we generated three tables, as shown in **Figure 3.** In the table ***donor***, each row represents an individual given a unique identification number (Donor ID). Other fields provided information about the clinical study in which an individual was enrolled (Study ID and Study Internal ID), gender, race, gestational age, whether the individual was breastfed and for how long, and whether the individual was enrolled in day care and if so, starting at what age. The second table, named ***donor_visits*** describes information about the donor’s age, CMV status, Epstein-Barr virus (EBV) status, Body Mass Index (BMI) and vaccine received on each clinical visit. Each clinical visit was given a unique identification (visit ID) in addition to the internal visit ID to distinguish between visits in consecutive years. For each visit, we calculated vaccine response by measuring HAI antibody response. Information about vaccine outcome is available as geometric mean (geo_mean), difference in the geometric mean before and after vaccination (delta_geo_mean), and difference for each vaccine strain (delta_single). Finally, each individual is classified as high and low responder (vaccine_resp). On each visit, samples were analysed and

information about which assays were performed and value of the measured analytes are stored in the ***experimental_data*** table. The FluPRINT database is available for download as **Data citation 3.**

Code availability

The source code for the PHP script is available from a public github repository (https://github.com/LogIN-/fluprint). Raw data files used to generate dataset are provided as **Data citation 1**. Full study population with demographic characteristics is provided as single CSV file as **Data citation 2**. The FluPRINT database is available from a research data repository Zenodo (**Data citation 3**). All files and source code are also available at the project website https://fluprint.com.

Data Records

The FluPRINT dataset described herein is available online for use by the research community and can be downloaded directly from a research data repository Zenodo (**Data citation 3)**. The dataset can be imported in the MySQL database for further manipulation and data extraction. The instructions how to import FluPRINT into the database are available at github (https://github.com/LogIN-/fluprint). The summary of the dataset, including the number of observations, fields and description for each table is provided in **Table 3**.

Technical Validation

The FluPRINT dataset was validated on two levels: (1) upon insertion and (2) after the data was inserted into the database. To validate data on insertion, we created loggers to monitor import of the CSV files into the database. This ensured easier and more effective troubleshooting of potential problems and contributed to the monitoring of the import process. Two different sets were used: (1) informative and (2) error loggers. Informative loggers provided information about which processing step has started or finished and how many samples have been processed in that particular step. This allowed us to monitor that correct number of samples was processed. Error loggers provided exact identification and name of the data which could not be imported into the database, usually caused by missing or incorrect user input, such as "…assay is missing. Skipping …'$row'". This facilitated the process to identify erroneous data, which were then manually reviewed, corrected, and updated.

Once the database was built, a manual review of data was performed to ensure accuracy and integrity of the dataset. Several random individuals were chosen and the accuracy of data was evaluated by comparison with the raw data. Additionally, we

evaluated total number of all donors, assays performed, clinical studies and years with the raw data available at the SDM.

Usage Notes

Recent advances in the computational biology and the development of novel machine learning algorithms, especially deep learning, make it possible to extract knowledge and identify patterns in an unbiased manner from large clinical datasets. Application of machine learning algorithms to clinical datasets can reveal biomarkers for different diseases, therapies[26], including vaccinations[2,3,6]. The data from the FluPRINT study can be used to identify cellular and molecular baseline biomarkers that govern successful antibody response to influenza vaccines (inactivated influenza vaccine, IIV and LAIV) across different influenza seasons and a broad age population. Since FluPRINT dataset is provided as a database, this facilitates further analysis. Queries can be easily performed to obtain a single CSV file. For example, researchers interested in understanding which immune cells and chemokines can differentiate between high and low responders that received inactivated influenza vaccine could search the FluPRINT database. In the database, they can find all donors for which flow cytometry or mass cytometry were performed together with Luminex assays, for which donors the HAI response was measured, and all the donors who received inactivated influenza vaccine. The resulting CSV file can then easily be used for downstream analysis.

Major advantages of this dataset are the mapping of the vaccine outcome, classifying individuals as high or low responders, standardization of the data from different clinical studies, and from different assays. This data harmonization process allows for direct comparison of immune cell frequency, phenotype, and functionality and quantity of chemokines and cytokines shared between individuals before or after influenza vaccinations. By releasing the FluPRINT database and the source code, we provide users with the ability to continue building upon this resource and to update the database with their data and other databases.

Acknowledgements

We are grateful to all individuals that participated in the research studies. We appreciate helpful discussions from all members of the Davis and Y. Chien labs. We also thank all staff members from the HIMC (Yael Rosenberg-Hasson and Michael D. Leipold) for data analysis, management and helpful discussions and HIMC Biobank (Rohit Gupta and Janine Bodea Sung) for sample processing and storage, Stanford-LPCH Vaccine Program (Alison Holzer) for management of clinical studies. This work

was supported by NIH grants (U19 AI090019, U19 AI057229) and the Howard Hughes Medical Institute to M.M.D, and by the EU's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant (FluPRINT, Project No 796636) to A.T.

Author contributions

A.T. downloaded and processed raw data, coordinated the integration of the data into the FluPRINT database, advised on the database design and wrote the manuscript. I.T. built the MySQL database, wrote PHP script for the data import into the database, and contributed to writing the manuscript. H.T.M. managed the analysis, data collection and management of the SDM at the HIMC. C.L.D. was responsible for regulatory approvals, protocol design, study conduct, and clinical data management. M.M.D. supervised the study.

Competing interests

The authors declare that they have no conflict of interests.

Figures

Figure 1. Overview of the FluPRINT dataset.

Figure 2. Demographic characteristics for the FluPRINT study population stratified by the vaccination outcome.

Figure 3. The FluPRINT database model.

Figure Legends

Figure 1. Overview of the FluPRINT dataset. The FluPRINT dataset consists of the 747 individuals from 8 clinical studies (SLVP015, SLVP017, SLVP018, SLVP021, SLVO024, SLVP028, SLVP029 and SLVP030) and 8 influenza seasons (from 2017 to 2015). Pie chart shows distribution of donors across clinical studies. The dataset contains harmonized data from different assays, including mass and flow cytometry, phosphorylated cytometry (Phospho-flow), multiplex ELISA (Luminex assay), clinical lab tests, such as complete blood test, analysis of hormones and virological assays (CMV and EBV antibody titers) and serological profiling with hemagglutination inhibition assay, which was used to define high and low responders.

Figure 2. Demographic characteristics for the FluPRINT study population stratified by the vaccination outcome. Distribution of individuals in the categories of high (red, n = 111) and low (grey, n = 253) responders regarding the (**a**) gender,

ethnicity and CMV status (**b**) age distribution between high and low responders. Age is indicated in years.

Figure 3. The FluPRINT database model. The diagram shows a schema of the FluPRINT database. Core tables, donors (red), donor_visits (yellow) and experimental_data (blue) are interconnected, while table donor_vaccines (orange) is connected to core table donors. The data fields for each table are listed, including the name and the type of the data. CHAR and VARCHAR, string data as characters; INT, numeric data as integers; FLOAT, approximate numeric data values; DECIMAL, exact numeric data values; DATETIME, temporal data values. Maximal number of characters allowed in the data fields is denoted as number in parenthesis.

Tables

Table 1. Demographic characteristics for the FluPRINT study population.

Age (y)	
Mean ± SD	38 ± 25
Median (range)	27 (1-90)
Gender	
Male (%)	295 (40%)
Female (%)	452 (60%)
Ethnicity	
European American (%)	497 (66.5%)
African American (%)	13 (1.8%)
American Indian and Alaska Native (%)	3 (0.4%)
Asian (%)	87 (11.6%)
Hispanic or Latino (%)	5 (0.7%)
Other (%)	137 (18.3%)
Unknown (%)	5 (0.7%)

Table 2. All assays performed across years in influenza clinical studies.

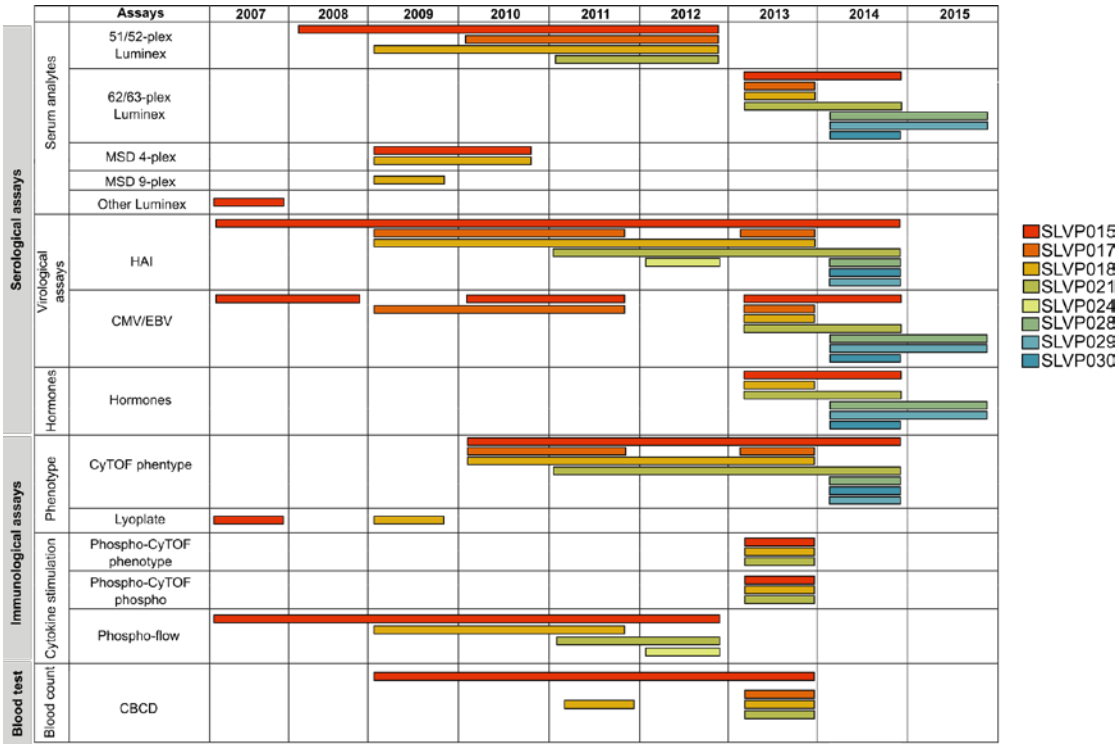


Table 3. The characteristics of the FluPRINT database.

Table name	Rows	Columns	Description
donors	820	11	Each row in this table is one donor. Donor is described with 11 additional parameters that include unique identification (*donor_id* and *study_donor_id*), identification for the study (*study_id*), full internal name of the study (*study_internal_id*), gender, race, gestational age (*age_birth*), information if donor was breastfed and for how long (*breastfed_period*) and information if donor was in day care and from which age (*daycare_age*).
donor_visits	2,944	18	Each row represents a donor at the particular visit (*visit_id*). Additionally, information about internal visit identification (*visit_internal_id*), date of the visit (*visit_year* and *visit_day*), pre- or post-vaccination visit for HAI assay (*visit_type_hai*), age at the visit (*age* and *age_round*), CMV/EBV status, BMI index at the visit are provided. Additionally, type of vaccine received (*vaccine*) and other calculated measures for HAI assay (*geo_mean, d_geo_mean, d_single, vaccine_resp*) are provided.
experimental_data	371,260	9	Each row represents a donor at particular visit (*donor_visits_id*). At each visit, assay that was performed is listed (*assay)* along with the names and values for measured analytes (*name, name_formatted, subset, units* and *data*).

References

1 Iuliano, A. D. *et al.* Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. *Lancet* **391**, 1285-1300, doi:10.1016/S0140-6736(17)33293-2 (2018).

2 Furman, D. *et al.* Systems analysis of sex differences reveals an immunosuppressive role for testosterone in the response to influenza vaccination. *Proc Natl Acad Sci U S A* **111**, 869-874, doi:10.1073/pnas.1321060111 (2014).

3 Furman, D. *et al.* Apoptosis and other immune biomarkers predict influenza vaccine responsiveness. *Mol Syst Biol* **9**, 659, doi:10.1038/msb.2013.15 (2013).

4 Furman, D. *et al.* Cytomegalovirus infection enhances the immune response to influenza. *Sci Transl Med* **7**, 281ra243, doi:10.1126/scitranslmed.aaa2293 (2015).

5 Nakaya, H. 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:10.1016/j.immuni.2015.11.012 (2015).

6 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).

7 Sobolev, O. *et al.* Adjuvanted influenza-H1N1 vaccination reveals lymphoid signatures of age-dependent early responses and of clinical adverse events. *Nat Immunol* **17**, 204-213, doi:10.1038/ni.3328 (2016).

8 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).

9 Tsang, J. S. *et al.* Global analyses of human immune variation reveal baseline predictors of postvaccination responses. *Cell* **157**, 499-513, doi:10.1016/j.cell.2014.03.031 (2014).

10 Sridhar, S. *et al.* Cellular immune correlates of protection against symptomatic pandemic influenza. *Nat Med* **19**, 1305-1312, doi:10.1038/nm.3350 (2013).

11 Bentebibel, S. E. *et al.* Induction of ICOS+CXCR3+CXCR5+ TH cells correlates with antibody responses to influenza vaccination. *Sci Transl Med* **5**, 176ra132, doi:10.1126/scitranslmed.3005191 (2013).

12 Trieu, M. C. *et al.* Long-term Maintenance of the Influenza-Specific Cross-Reactive Memory CD4+ T-Cell Responses Following Repeated Annual Influenza Vaccination. *J Infect Dis* **215**, 740-749, doi:10.1093/infdis/jiw619 (2017).

13 Hagan, T., Nakaya, H. I., Subramaniam, S. & Pulendran, B. Systems vaccinology: Enabling rational vaccine design with systems biological approaches. *Vaccine* **33**, 5294-5301, doi:10.1016/j.vaccine.2015.03.072 (2015).

14 Chattopadhyay, P. K., Gierahn, T. M., Roederer, M. & Love, J. C. Single-cell technologies for monitoring immune systems. *Nat Immunol* **15**, 128-135, doi:10.1038/ni.2796 (2014).

15 Galli, E. *et al.* The end of omics? High dimensional single cell analysis in precision medicine. *Eur J Immunol*, doi:10.1002/eji.201847758 (2019).

16 Bendall, S. C., Nolan, G. P., Roederer, M. & Chattopadhyay, P. K. A deep profiler's guide to cytometry. *Trends Immunol* **33**, 323-332, doi:10.1016/j.it.2012.02.010 (2012).

17 Simoni, Y., Chng, M. H. Y., Li, S., Fehlings, M. & Newell, E. W. Mass cytometry: a powerful tool for dissecting the immune landscape. *Curr Opin Immunol* **51**, 187-196, doi:10.1016/j.coi.2018.03.023 (2018).

18 Newell, E. W., Sigal, N., Bendall, S. C., Nolan, G. P. & Davis, M. M. Cytometry by time-of-flight shows combinatorial cytokine expression and virus-specific cell niches within a continuum of CD8+ T cell phenotypes. *Immunity* **36**, 142-152, doi:10.1016/j.immuni.2012.01.002 (2012).

19 The Human Immune Monitoring Center. *Standard protocols of the Human Immune Monitoring Center*, <https://iti.stanford.edu/himc/protocols.html> (2018).

20 Siebert, J. C., Munsil, W., Rosenberg-Hasson, Y., Davis, M. M. & Maecker, H. T. The Stanford Data Miner: a novel approach for integrating and exploring heterogeneous immunological data. *J Transl Med* **10**, 62, doi:10.1186/1479-5876-10-62 (2012).

21 Brodin, P. *et al.* Variation in the human immune system is largely driven by non-heritable influences. *Cell* **160**, 37-47, doi:10.1016/j.cell.2014.12.020 (2015).

22 Whiting, C. C. *et al.* Large-Scale and Comprehensive Immune Profiling and Functional Analysis of Normal Human Aging. *PLoS One* **10**, e0133627, doi:10.1371/journal.pone.0133627 (2015).

23 Leipold, M. D. & Maecker, H. T. Phenotyping of Live Human PBMC using CyTOF Mass Cytometry. *Bio Protoc* **5** (2015).

24 Fernandez, R. & Maecker, H. Cytokine-stimulated Phosphoflow of PBMC Using CyTOF Mass Cytometry. *Bio Protoc* **5** (2015).

25 Hirst, G. K. The Quantitative Determination of Influenza Virus and Antibodies by Means of Red Cell Agglutination. *J Exp Med* **75**, 49-64 (1942).

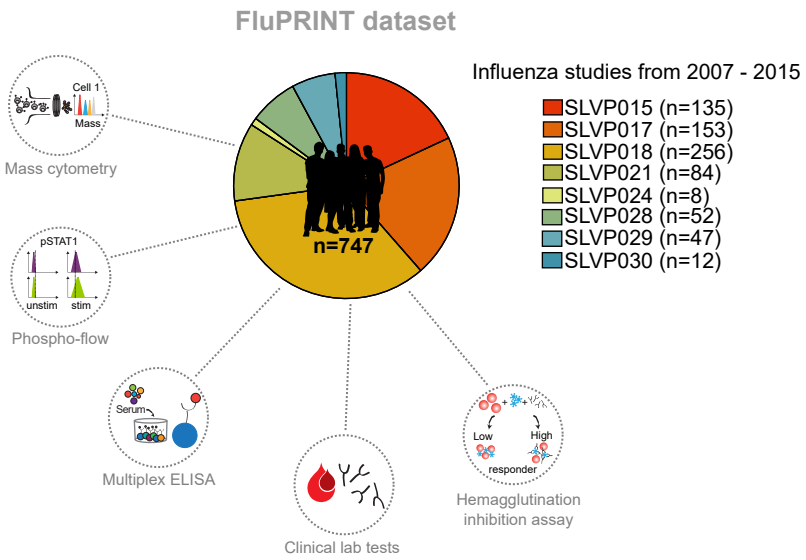
26 Krieg, C. *et al.* High-dimensional single-cell analysis predicts response to anti-PD-1 immunotherapy. *Nat Med* **24**, 144-153, doi:10.1038/nm.4466 (2018).

Data Citations

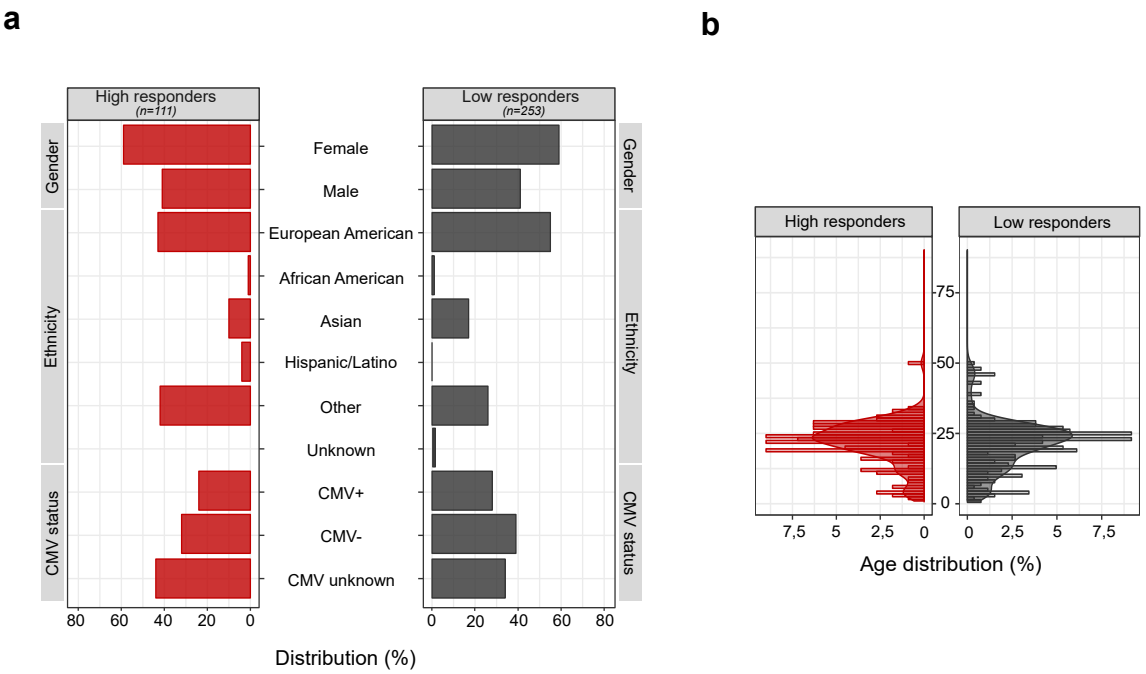
1. Tomic, A. and Tomic, I. Zenodo https://doi.org/10.5281/zenodo.2546003 (2019).

2. Tomic, A. and Tomic, I. Zenodo https://doi.org/10.5281/zenodo.2546010 (2019).

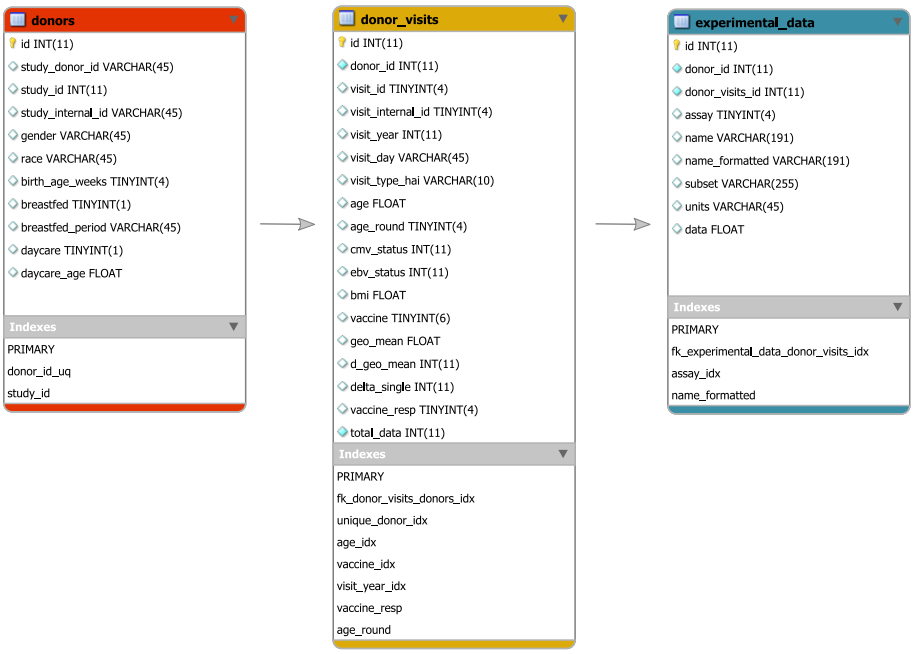
3. Tomic, A. and Tomic, I. Zenodo https://doi.org/10.5281/zenodo.2546017 (2019).



Tomic et al.
Figure 1.



Tomic et al.
Figure 2.



Tomic et al.
Figure 3.