Title
The 10,000 Immunomes Project: Building a Resource for Human Immunology

Journal
Cell Reports, 25(2)

ISSN
2639-1856

Authors
Zalocusky, Kelly A
Kan, Matthew J
Hu, Zicheng
et al.

Publication Date
2018-10-01

DOI
10.1016/j.celrep.2018.09.021

Peer reviewed

HHS Public Access

Author manuscript

Cell Rep. Author manuscript; available in PMC 2018 November 29.

Published in final edited form as:

Cell Rep. 2018 October 09; 25(2): 513–522.e3. doi:10.1016/j.celrep.2018.09.021.

The 10,000 Immunomes Project: Building a Resource for Human Immunology

Kelly A. Zalocusky[1,2], **Matthew J. Kan**[1,2], **Zicheng Hu**[1,2], **Patrick Dunn**[3], **Elizabeth Thomson**[3], **Jeffrey Wiser**[3], **Sanchita Bhattacharya**[1,2,4], and **Atul J. Butte**[1,2,4,5,*]

[1]Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA 94158, USA

[2]Department of Pediatrics, University of California, San Francisco, San Francisco, CA 94158, USA

[3]Information Systems Health IT, Northrop Grumman, Rockville, MD 20850, USA

[4]Senior Author

[5]Lead Contact

SUMMARY

There is increasing appreciation that the immune system plays critical roles not only in the traditional domains of infection and inflammation but also in many areas of biology, including tumorigenesis, metabolism, and even neurobiology. However, one of the major barriers for understanding human immunological mechanisms is that immune assays have not been reproducibly characterized for a sufficiently large and diverse healthy human cohort. Here, we present the 10,000 Immunomes Project (10KIP), a framework for growing a diverse human immunology reference, from ImmPort, a publicly available resource of subject-level immunology data. Although some measurement types are sparse in the presently deposited ImmPort database, the extant data allow for a diversity of robust comparisons. Using 10KIP, we describe variations in serum cytokines and leukocytes by age, race, and sex; define a baseline cell-cytokine network; and describe immunologic changes in pregnancy. All data in the resource are available for visualization and download at http://10kimmunomes.org/.

*Correspondence: atul.butte@ucsf.edu,.

AUTHOR CONTRIBUTIONS

K.A.Z. led the design and implementation of the work, in collaboration with Z.H. for analysis of cytometry data, with P.D., E.T., and J.W. for building and curation of the ImmPort database, and with M.J.K., S.B., and A.J.B. for continual feedback regarding design and analysis throughout the project. The paper was written by K.A.Z. with editorial input from all authors. All data are available at http://10kimmunomes.org/.

DECLARATION OF INTERESTS

The authors declare no competing interests.

DATA AND SOFTWARE AVAILABILITY

All raw data used for this study are available through the ImmPort Data Portal: https://aspera-immport.niaid.nih.gov:9443/login. Every study accessed is enumerated in Table S1. Formatted and normalized data tables are available for download at http://10kimmunomes.org

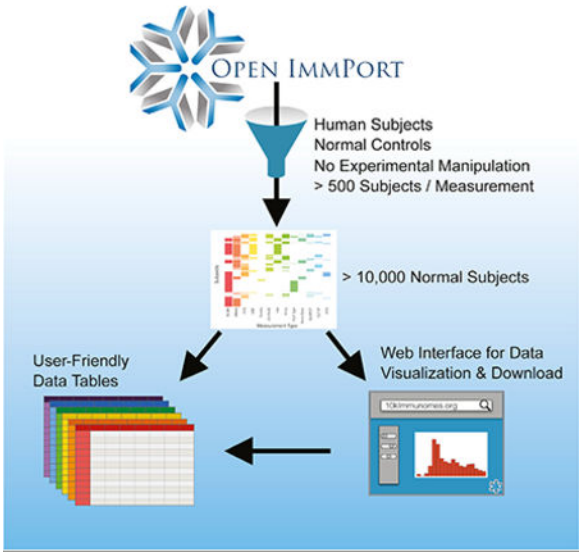
SUPPLEMENTAL INFORMATION

Supplemental Information includes six figures and one table and can be found with this article online at https://doi.org/10.1016/j.celrep.2018.09.021.

In Brief

Zalocusky et al. report the development of a data resource comprising curated, integrated, and normalized immunology measurements from all healthy normal human subjects in the ImmPort database

Graphical Abstract



INTRODUCTION

The advancement of technologies in preclinical immunology (Elshal and McCoy, 2006; Leng et al., 2008; Maecker et al., 2010; Saeys et al., 2016; Spitzer and Nolan, 2016) and the promise of precision therapeutics in immunology (Ashley, 2015; Collins and Varmus, 2015; Friedman et al., 2015), have together propelled a rapid increase in the production of large-scale immunological data. Similar advancements in other fields, such as genomics, where high-throughput assays spurred a swell of data, have demonstrated the need and benefit of common reference datasets. Resources such as the 1000 Genomes Project (1000 Genomes Project Consortium, 2010, 2012; Sudmant et al., 2015), Health and Retirement Study (https://hrs.isr.umich.edu/), Wellcome Trust Case Control Consortium (Burton et al., 2007), and Exome Aggregation Consortium (Lek et al., 2016) have accelerated discovery of thousands of disease-linked variants and uniquely enable understanding of global variation in the human genome in health and disease. To date, however, human immunology has no such resource. As publicly available immunological data continue to grow, there is a need in the field for a framework in which those data can be accumulated and normalized, allowing researchers to explore the space of existing data and generate testable hypotheses.

The challenge in generating such a resource lies, in part, in the diversity of data types available to immunologists. A reference “immunome” might reasonably include flow cytometry, gene expression, human leukocyte antigen (HLA) type, cytokine measurements,

clinical assessments, and more. Furthermore, standardized protocols for measurement and conventions for naming cell types and cytokines are only currently being developed, and adherence is inconsistent (Finak et al., 2016). For the experimental or clinical immunologist, the cost of generating the necessary data from scratch—or the temporal and computational costs associated with standardizing and harmonizing data from publicly available cohorts across platforms, time points, and institutions—is prohibitive without significant resources. Thus, although the benefit of a common reference population is clear, and large-scale data are publicly available, this need has not yet been met.

Other lessons from the field of genomics offer additional direction and promise. For example, resources like the 1000 Genomes Project (1000 Genomes Project Consortium, 2010, 2012; Sudmant et al., 2015) have clearly demonstrated the necessity of exploring and accounting for human diversity; the publication of the original data release has been cited more than 5,000 times. Additionally, although high-throughput assays invariably suffer from inter-experiment technical variation, the field has generated and validated statistical methods for overcoming those artifacts while preserving the underlying effects of interest (Hughey and Butte, 2015; Johnson et al., 2007; Leek and Storey, 2007; Leek et al., 2010; Pelz et al., 2008). These breakthroughs, ripe for translation to immunological data, have unlocked the potential for deeper insight beyond the initial intent of each of the thousands of studies that have made their raw data publicly available to researchers.

Given the recent growth in open immunology data, we sought to establish a structure for synthetically constructing a reference “immunome” by integrating individual-level data from publicly available immunology studies. For this initial version, we began by manually curating the entire public contents of ImmPort (Data Release 21; http://www.immport.org/), the archival basic and clinical data repository and analysis platform for the National Institute for Allergy and Infectious Disease (NIAID) (Bhattacharya et al., 2014; Dunn et al., 2015). ImmPort contains studies on a diversity of topics related to immunity, including allergy, transplant, vaccinology, and autoimmune disease, and the data represented are diverse, ranging from flow cytometry and ELISA to clinical lab tests and HLA type. While most of these studies were not designed to examine the diversity of the healthy normal immune system, they nonetheless contain healthy control arms that we utilized for this purpose.

Our goal was to include in the reference only human subjects from the healthy control arms of studies and only samples from individuals that have undergone no experimental manipulation. Our filtering and data harmonization process resulted in an inaugural dataset consisting of 10 data types in standardized tables (mass cytometry [cytometry by time of flight; CyTOF], flow cytometry, multiplex ELISA, gene expression array, clinical lab tests, and others) on samples taken from 10,344 subjects. We unify the data from all the normal healthy immunomes into a fully open and interactive online resource (http://10kimmunomes.org/), which to date has accumulated >4,200 distinct users. We expect that the ability to dynamically visualize the reference will accelerate discovery in immunology. We further show that this resource can provide a basis for studying immunity across age, sex, and racially diverse populations. The ImmPort Data Curation team is supporting the maintenance of the project as an arm of the ImmPort environment (http://www.immport.org/resources), ensuring that the 10,000 Immunomes Project (10KIP) will only grow in value,

richness, and scale with the participation of the immunology community in the open-data movement

RESULTS

Development of the 10KIP

To develop the 10KIP, we began with ImmPort Data Release 21 (downloaded May 3, 2017), which contains 242 studies released to the public, with 44,775 subjects and 293,971 samples (Figure 1). We began by manually curating each of these 242 studies, reading inclusion and exclusion criteria, and selecting by hand which study arms and planned visits constitute data collected on samples from normal healthy human subjects prior to any experimental immune perturbation. This manual curation process resulted in an inaugural population of 10,344 subjects, spanning 83 studies. An exhaustive list of all studies, arms, and planned visits that qualified for inclusion is available as Table S1.

This dataset consists of 10 distinct data types (flow cytometry, high-throughput serum protein measurements, gene expression, clinical lab tests, and others). For each of the 10 data types, we developed a standardized pipeline for data cleaning and harmonization (see STAR Methods). Although the ImmPort environment boasts an exceptional degree of annotation, massively multi-study analysis still required substantial effort in data harmonization. Across all studies, we standardized analyte names and units of measurement, segregated data by sample type (e.g., peripheral blood mononuclear cells [PBMCs] versus whole blood versus serum), and corrected for differences in sample dilutions. This process resulted in standardized data tables, which form the backbone of the reference. The normalized data and their raw counterparts are available for visualization and download at http://10kimmunomes.org/ (Figure S6).

Contents of the Inaugural Release of the 10KIP

The initial release of the 10KIP contains 10,344 subjects. They are approximately evenly split between male and female, represent a diverse racial makeup, and include more than 1,000 pediatric subjects (<18 years of age) and over 1,300 subjects above 65 years of age (Figure S1). As enumerated in Table 1, the resource contains secreted protein data from over 4,800 subjects, clinical lab test data from over 2,600 subjects, flow cytometry or mass cytometry data from over 1,400 subjects, hemagglutination inhibition (HAI) titers from over 1,300 subjects, and HLA types from over 1,000 subjects, in addition to several other data types. Because many subjects contribute more than one type of measurement, the total number of subjects across all measurement types substantially exceeds the number of distinct subjects. Data are available in “formatted” or “normalized” formats. “Formatted” data are segregated by biological sample type and harmonized to include standardized units of measurement and analyte names, but these data are not batch corrected. “Normalized” files are the batch-corrected versions of the formatted data (STAR Methods). Individuals wishing to verify the batch correction may view a whole-data visualization of each data type (Figure S5), view the effect of study on every analyte individually on http://10kimmunomes.org, or to download the formatted data and conduct their preferred method of batch correction.

Multiplex ELISA Measurements across the Population

The regulation of immune system components through cytokines, chemokines, adhesion molecules, and growth factors is central to maintenance of a healthy immune homeostasis and response to acute infection (Becher et al., 2017; Mackay, 2001; Neurath, 2014; O'Shea et al., 2002). Recent advances in the measurement of such secreted proteins with multiplex ELISA (also known as multiplex bead-based analysis or by the trade name Luminex) allow for high-throughput profiling of the immune molecular milieu (Morgan et al., 2004; Vignali, 2000). Similar to high-throughput measurements of RNA expression, however, this type of measurement must be interpreted with caution, due to inter-experimental technical variation, as well as differences in reagents and platforms used (Djoba Siawaya et al., 2008). In fact, there is contention within the field of computational immunology regarding the validity of directly comparing high-throughput serum cytokine measurements across studies (Khan et al., 2004; Rosenberg-Hasson et al., 2014).

Here we suggest, however, that previously described models for statistical compensation for batch effects in genomics are sufficient for analysis of multiplex ELISA data. We find that, without batch correction, technical variation contributes significantly to the clustering of multiplex ELISA data as visualized by t-SNE (Figure 2A). The empirical Bayes algorithm ComBat (Johnson et al., 2007), originally designed for analysis of microarray data, compensates for both mean and variance differences across studies while preserving potential effects of interest, such as differences by age, sex or race (Figures 2B and S2). We have additionally confirmed the efficacy of this strategy through 1,000-fold simulations of multiplex ELISA data with mean, variance, and single-analyte batch effects (Figure S2). This strategy preserves known effects, such as a significantly higher serum leptin concentration in women as compared to men (Figure 2C; Dubuc et al., 1998).

Additionally, the ability to combine data across studies from disparate geographic locations and distinct ethnic populations enables us to uncover demographic diversity in cytokine and chemokine expression. We have systematically analyzed the multiplex ELISA data and report all significant associations after accounting for multiple comparisons (Figure S3). Strikingly, we find that 10 out of the 50 most commonly measured cytokines, chemokines, and metabolic factors measured by multiplex ELISA differ significantly by race (Figures 2 and S3). Nineteen differ significantly with age and 3 by sex. To highlight one example, our analyses indicate a significantly higher level of C-X-C motif chemokine 5 (CXCL5) among African Americans as compared to other races (Figure 2D). Interestingly, CXCL5 has been implicated in a number of disorders that disproportionately affect African Americans, including acute coronary syndrome (Sirak et al., 2008; Zineh et al., 2008), chronic obstructive pulmonary disease (Kirkpatrick and Dransfield, 2009; Qiu et al., 2003), asthma (Joseph et al., 2000; Qiu et al., 2007), and insulin resistance (Chavey et al., 2009; Spanakis and Golden, 2013). Finally, this analysis from a population of up to 1,286 individuals across 17 studies (STAR Methods) allows us to describe the distribution of serum cytokine measurements in a diverse human population (Figure 2E). Some, such as interleukin-5 (IL-5) and IL-7, lie within a relatively small range, whereas others, such as chemokines C-C motif chemokine 4 (CCL4) and (CXCL9), display a many-fold range, even within this population of putative healthy normal human subjects. Together, these findings affirm the

benefit of maintaining and growing a diverse common control population for the future of clinical and precision immunology.

Individual-Level Cell-Subset Measurements across the Population

Similarly, even within this reference population, we find a high degree of variability in the proportion of immune cell subsets from PBMCs as measured by mass cytometry. This variability in cell subsets within a normal healthy population corroborates previously reported descriptions of cell-subset percentages (Brodin and Davis, 2017). CD4+, CD8+, and gamma-delta T cell subsets, in particular, span a wide range as a percentage of total leukocytes. Smaller subsets, such as memory B cells and plasmablasts, span a tighter range (Figure 3A). For high-throughput analysis of mass cytometry data, we have employed a pipeline (MetaCyto) that begins with raw flow cytometry standard (fcs) files, enacts quality control, implements automated gating based on a standard set of markers, and reports a standardized set of cell-subset percentages as a proportion of total leukocytes (STAR Methods; Hu et al., 2017). In a prior publication, we utilized this pipeline to enumerate a number of associations between race and cell-subset percentages from analysis of publicly available data (Hu et al., 2017).

Here, we describe the effects of age (Figures 3B and 3C), sex (Figures 3D and 3E), and race (Figure S4) in this larger healthy normal population. In total, we find that four of the 24 measured cell types differ significantly by race (Figure S4), 20 change significantly with age (Figures 3B and 3C), and 7 vary significantly by sex (Figures 3D and 3E). As an example, our analysis reveals a pronounced decline in naive CD8+ T cells with age (Figure 3C) with a concomitant increase in memory CD4+ T cells (Figure 3C). These findings are anticipated given the accumulation of antigen exposures over the lifespan. Our analysis additionally suggests that women have significantly higher levels of naive $CD4^+$ T cells, naive CD8+ T cells, naive B cells, and plasmablasts than male subjects while having a significantly smaller proportion of effector CD8+ T cells (Figures 3D and 3E). We additionally find that natural killer (NK) cells are found at a significantly higher level in Asian subjects than white subjects and that regulatory T cells are measured at a significantly higher level in African American subjects as compared to all other races (Figure S4). These age, sex, and race-related differences in immune cell subsets may help explain population differences in infections and autoimmune disease or impact clinical decision-making as it pertains to treatment selection. Developing and continuing to grow a diverse reference of immune measurements specifically enables this type of discovery.

Systems-Level Network Analysis of Cellular and Molecular Immunity

In addition to characterizing the diversity of the immune system in terms of cellular and molecular markers, a framework such as the 10KIP also has the potential to facilitate systems-level network analysis. We selected 321 individuals from the dataset for which immune cell subsets in PBMCs and protein measurements of serum cytokines, as measured by mass cytometry and multiplex ELISA, respectively, were assessed in the same biological samples. We modeled the partial correlation between each cell type and each cytokine, statistically controlling for age, sex, and race (Figure 4), all of which our analyses suggest can have significant effects on cellular and molecular immune repertoire (Figures 2 and 3),

and then display only those correlations that remain significant at a false discovery rate (FDR) of 0.01 following Benjamini-Hochberg (BH) correction. Although we recognize the potential for generating a network that incorporates additional datatypes, such as gene expression, as an initial demonstration, we restrict ourselves to a bipartite network of cells and cytokines, generating a rich but interpretable network.

Our analysis recovers some known relationships; for example, we see that effector $CD4^+$ T cells function as a major hub in the network, contributing positive associations with known Th2 cytokines IL-5, IL-10, and IL-13 (Wynn, 2003). We additionally see a negative association between regulatory T (Treg) cells and the pro-inflammatory CSF3 (formerly granulocyte colony stimulating factor [GCSF]), consistent with the known immunomodulatory role of Treg cells (Belkaid and Rouse, 2005; Vignali et al., 2008). We detect an association between CXCL10 and monocyte subsets, concordant with evidence that this cytokine is expressed by and acts to recruit monocytes (Lee et al., 2009). Furthermore, acute phase reactants interferon alpha-2 and IL-6 are negatively associated with central memory CD8+ T cells and memory B cells, which is concordant with the understanding of the kinetics of the transition from acute inflammation to memory formation (Huber and Farrar, 2011). This exploratory analysis of the cell-cytokine network in the normal, healthy immune system also generates testable hypotheses about human immune function. For example, this analysis suggests a positive association between leptin and transitional and memory B cells, connections that are potentially of interest given B cell expression of the leptin receptor and the recent discovery that B cells may promote insulin resistance (La Cava and Matarese, 2004; Winer et al., 2011). Furthermore, this connection through memory B cells extends to the adipokine resistin and to the adhesion molecules ICAM-1 and VCAM-1, a cluster of molecules also known to be affected by adiposity (Procaccini et al., 2013; Skilton et al., 2005; Verma et al., 2003). These analyses together demonstrate the utility of this framework for generating systems-level hypotheses from publicly available immunology data collected for a variety of disparate purposes.

Use as a Common Control Population for Precision Immunology in Pregnancy

Finally, to illustrate the potential of the 10KIP to serve as a common control group for clinical studies, we used an age- and sex-matched subset of the 10KIPto compare with immune measurements in pregnancy, derived from ImmPort study SDY36. In this ImmPort study, researchers collected rich clinical data, as well as flow cytometry and serum cytokine measurements, from a population of 56 women during each trimester of pregnancy, 6 weeks postpartum, and 6 months postpartum. Cell count data from this study, as well as trends in cytokine secretion from cultured cells, have been published previously (Kraus et al., 2012). Changes in serum cytokine levels over gestation and analyses of cell-subset percentages (which are potentially differentially affected during pregnancy; Ekouevi et al., 2007), however, remain undescribed as of this writing. Additionally, the study design did not incorporate a pre-pregnancy control, leaving open the question of whether cell subsets and serum cytokines truly return to baseline by 6 months postpartum. Given work demonstrating persistence of fetal cells and DNA in maternal blood and brain many years postpartum (Bianchi et al., 1996; Chan et al., 2012), the comparison to a common healthy control has the potential to enrich our understanding of the immune system in pregnancy and maternity.

We first applied principal-component analysis (PCA) to the serum cytokine measurements, which revealed a major shift in cytokine regulation during the first trimester of pregnancy as compared to second and third trimester measurements, postpartum measurements, and measurements taken from age and sex-matched 10KIP controls (Figure 5A). This shift is primarily driven by increased concentrations of CCL2, CCL3, CCL4, CCL5, CCL11, CXCL10, and IL-6. As an example of this modulation, we see that CCL5 concentration is significantly increased during the first and second trimester and is decreased during the third trimester and up to 6 weeks postpartum, but it returns to baseline by 6 months postpartum (Figure 5B). In contrast, IL-15 measurements remain relatively constant over the entire course of gestation (Figure 5C).

In addition to analysis of serum cytokine concentrations, we also examined changes in cell-subset percentages in pregnancy. PCA analysis of flow cytometry measurements indicated that changes in cell subsets over the course of gestation are not the primary source of variation as compared to postpartum or reference measurements (Figure 5D). This is not to say, however, that cell subsets remain static over the course of pregnancy. We see, for example, that CD4+ T cells, as a percent of lymphocytes, undergo a significant increase during all three trimesters of pregnancy as compared to the 10KIP reference population (Figure 5E). B cells, in contrast, exhibit a small but significant dip during the second and third trimesters (Figure 5F). Although we recognize that, to date, ImmPort does not contain dense data for all measurement types, this analysis demonstrates that the size and scope of even this initial version of the 10KIP are sufficient to generate age- and sex-matched control cohorts for two types of high-throughput immune measurements as a baseline or comparator to immune perturbation or disease.

DISCUSSION

Although the availability of large common control cohorts, such as the 1000 Genomes Project (1000 Genomes Project Consortium, 2010, 2012; Sudmant et al., 2015) and the Wellcome Trust Case Control Consortium (Burton et al., 2007), has proven immensely useful for various biological research communities, no parallel resource exists for immunological measurements. Here, we produced, through manual curation and study-by-study harmonization, the 10,000 Immunomes Project (10KIP), a framework for growing a standardized reference dataset for the immunology community. To enable its use by experimental and clinical immunologists, we developed an interface for interactive data visualization, as well as custom cohort creation and data download (available at http://10kimmunomes.org/). Through statistical testing and validations in simulated data, we demonstrate the ability to compensate for technical artifacts that invariably arise from collecting data on different days, across different platforms, or at distant institutions by repurposing algorithms developed in computational genetics.

We not only recover many known differences by age and sex across serum cytokine and cell-subset measurements but also reveal differences, particularly by race, that would have been impossible to uncover without the combination of dozens of independent datasets generated to answer varied and unrelated questions in immunology. Through network analysis, we additionally demonstrate the utility of the resource for generating insights into cell-cytokine

relationships in the human immune system. Finally, we demonstrate that the size and scope of the extant publicly available data are sufficient for custom cohort selection, enabling us to generate a reference cohort of women between the ages of 18 and 40 years who have both cell-subset and serum cytokine data available on the same blood samples for comparison with an external dataset derived from measurements taken during pregnancy. Generating a sufficiently powered sex- and age-matched population with multiple immune measurements for comparison allowed us to explore the cell-subset and cytokine changes that occur as the immune system is modulated over the course of gestation. We expect that as more and more measurements are uploaded to NIAID's ImmPort database, the 10KIP will expand into an increasingly powerful resource for future clinical and preclinical immunological studies.

While we recognize the ideal would be to recruit and collect immune measurements from a large cohort, the resources and time required to collect immunologic measurements from a sufficiently sized heterogeneous population would be considerable, while this sizable volume of subject-level human immunology data are currently available to the research community. We also acknowledge further potential limitations in this work. For example, we are selecting subjects, standardizing labels and units, and otherwise curating the data with the best available information on these studies, but it is possible errors in the original data descriptions or labeling might persist. Also, we present the data after normalization and batch correction, but of course, we recognize that all of these source datasets were collected independently across institutions, technologies, and time. It is possible that our normalization efforts and assumptions might not hold true for every analysis in every study. On the other hand, the large inter-study variation (see Figure S5) supports the idea that the batch-corrected consensus may be more robust than the results of each study taken separately. We also note that, to date, some data types might not be measured densely enough to make reliable models that span all ages or races, and to date, racial information in ImmPort is acquired at a relatively coarse grain. Additionally, some highly sought data types, such as RNA sequencing (RNA-seq), are not yet available in sufficient volume in ImmPort to merit inclusion in the initial release of the resource. As high-throughput immunological techniques become more widely available and as experimentalists continue to deposit these data in ImmPort, however, this scaffold will continue to grow, enabling well-powered analyses on more specific populations and over an increasing number of data types with time.

Finally, we want to recognize current reference datasets for immunology. The extant resources, while of clear import to the research community, serve different purposes than does the 10KIP. ImmGen (Shay and Kang, 2013), for example, represents an immense resource of immune gene expression in murine models, while the 10KIP instead focuses on multiple data types in human immunology. Likewise, ImmuneSpace (Sauteraud et al., 2016) provides a suite of visualization and analysis tools, allowing users to interact and download data at the level of individual human immunology studies. The 10KIP, in contrast, has as its primary goals to filter the extant data for only healthy normal subjects and enable visualization and analysis across many studies. The 10KIP takes full advantage of the structure of ImmPort, in which subjects are assigned a unique accession number and are associated with their age, sex, and race. The resource allows researchers to subset the population or to look for associations with these general demographic phenotypes.

Additionally, it leverages the richness of data available through ImmPort, which encompasses soluble protein and cytokine measurements, such as multiplex ELISA, cell-phenotyping measurements such as flow-cytometry and CyTOF, standard medical laboratory test panels, gene expression data, and others. We believe that integrating these datasets and presenting them as a fully open resource will pay dividends in terms of both basic research and the precision and robustness of ongoing translational efforts in immunology.

STAR⋆METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Studies from ImmPort Data Release 21	Detailed in Table S1	https://aspera-immport.niaid.nih.gov:9443/login
Software and Algorithms		
MetaCyto	(Huetal., 2017)	https://bioconductor.org/packages/release/bioc/html/MetaCyto.html
ComBat	(Johnson et al., 2007)	https://bioconductor.org/packages/release/bioc/html/sva.html
affy	(Gautier et al., 2004)	https://www.bioconductor.org/packages/devel/bioc/html/affy.html
GEOquery	(Davis and Meltzer, 2007)	https://bioconductor.org/packages/release/bioc/html/GEOquery.html
preprocessCore	(Bolstad, 2017)	https://www.bioconductor.org/packages/release/bioc/html/preprocessCore.html
impute	(Hastie et al., 2017)	https://bioconductor.org/packages/release/bioc/html/impute.html
RmySQL	(Ooms et al., 2018)	https://cran.r-project.org/web/packages/RMySQL/index.html
limma	(Ritchie et al., 2015)	https://bioconductor.org/packages/release/bioc/html/limma.html

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Atul J. Butte (atul.butte@ucsf.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Subjects were extracted from Data Release 21 of ImmPort database, which contains 242 open-access studies, together comprising 44,775 subjects and 293,971 samples. Each subject in ImmPort is assigned a unique identifier, allowing every measurement in the ImmPort database to be assigned to a unique subject. Each subject has, at minimum, race, age, and sex demographic information. The ImmPort data architecture requires that each study contain detailed descriptions of inclusion and exclusion criteria for subjects. Additionally, each arm (experimental and control arms) of each study is assigned a unique accession. Finally, each experimental measurement is time stamped with a unique planned visit accession. Manual review of the inclusion/exclusion criteria, arms, and planned visits allowed us to select control subjects, and to examine only those measurements taken before the onset of any experimental manipulation, such as vaccine, drug, or surgery that may have

occurred. A complete list of qualifying studies, arms, and planned visits contained in the 10KIP is available in Table S1.

METHOD DETAILS

Extract Immune Cell Frequencies from Cytometry Data—Meta-analysis of Cytometry data is conducted using the MetaCyto package (Hu et al., 2017). Briefly, flow cytometry data and CyTOF data of healthy human blood samples from ImmPort studies SDY89, SDY112, SDY113, SDY144, SDY167, SDY180, SDY202, SDY212, SDY296, SDY305, SDY311, SDY312, SDY314, SDY315, SDY364, SDY368, SDY387, SDY404, SDY420, SDY472, SDY475, SDY478, SDY514, SDY515, SDY519, SDY702, SDY720, and SDY736 were downloaded from ImmPort web portal. Flow cytometry data from ImmPort were compensated for fluorescence spillovers using the compensation matrix supplied in each fcs file. All data from ImmPort were arcsinh transformed. For flow cytometry data, the formula f(x) = arcsinh (x/150) was used. For CyTOF data, the formula f(x) = arcsinh (x/8) was used. Transformation and compensation were done using the *preprocessing.batch* function in MetaCyto (Hu et al., 2017). The cell definitions from the Human ImmunoPhenotyping Consortium (Finak et al., 2016) were used to identify 24 types of immune cells using the *searchClster.batch* function in MetaCyto. Specifically, each marker in each cytometry panels was bisected into positive and negative regions. Cells fulfilling the cell definitions are identified. For example, the CD14+ CD33+ $CD16^-$ ($CD16^-$ monocytes) cell subset corresponds to the cells that fall into the CD14+ region, CD33+ region and $CD16^-$ region concurrently. The proportion of each cell subsets in the PBMC or whole blood were then calculated by dividing the number of cells in a subset by the total number of cells in the blood. Differences by age, sex, and race were detected with a linear model, with Tukey's Honestly Significant Difference (Tukey's HSD) post hoc tests and Benjamini-Hochberg (BH) correction for false discovery rate.

Multiplex ELISA analysis—Secreted protein data measured on the multiplex ELISA platform were collected from ImmPort studies SDY22, SDY23, SDY111, SDY113, SDY180, SDY202, SDY305, SDY311, SDY312, SDY315, SDY420, SDY472, SDY478, SDY514, SDY515, SDY519, and SDY720. Data were drawn from the ImmPort parsed data tables using RMySQL or loaded into R from user-submitted unparsed data tables. Across the studies that contribute data, there are disparities in terms of the dilution of samples and units of measure in which the data are reported. We corrected for differences in dilution factor and units of measure across experiments and standardized labels associated with each protein as HUGO gene symbols. This step represents the "formatted" multiplex ELISA data table. For our own analysis, as represented in Figure 2, we analyzed only those proteins that were measured in more than half of the subjects leaving the 50 most-commonly measured proteins. Compensation for batch effects was conducted using the *ComBat* function of the R package sva, with study accession representing batch and a model matrix that included age, sex, and race of each subject. Data were log2 transformed before normalization with ComBat to better fit the assumption that the data are normally distributed (Figure S6). We verified that a linear model associating age, sex, ethnicity, and study accession of each subject no longer revealed any significant associations between study accession and protein concentration following batch correction, and that known differences, such as the difference

in leptin concentration by sex, were captured following our batch correction procedure. We additionally validated our approach using 1000-fold data simulations (see below). Differences by age, sex, and race were detected with a linear model, with Tukey's Honestly Significant Difference post hoc tests and Benjamini-Hochberg correction for false discovery rate. For dimensionality reduction analysis, which is not robust to missing values, missing values were imputed by k-nearest neighbors, using impute.knn with default values.

Network Analysis—The bipartite network depicted in Figure 4 represents an analysis over the 24 immune cell subset percentages calculated in the mass cytometry analysis described above in *Extract Immune Cell Frequencies from Cytometry Data* and the 50 soluble protein measurements, normalized and batch-corrected as described above in *Multiplex ELISA analysis.* Data were included from the 321 subjects where both multiplex ELISA and mass cytometry measurements were conducted on the same biological sample. Edges depict the Spearman's p of a partial correlation between each cytokine concentration and each individual cell type, accounting for age, sex, and race. Only correlations that remained significant at a BH-corrected $p < 0.01$ are shown.

Cell and cytokine modulation in pregnancy—We compared serum cytokine and cell subset percentages from 10KIP samples to measurements taken from women during and after pregnancy. We selected samples from the 10KIP from women aged 18–40 who contributed CyTOF data from PBMC and multiplex ELISA measurements. Samples from pregnancy were taken from ImmPort study SDY36. The serum cytokine and flow cytometry from SDY36 was batch corrected together with the ImmPort reference data, using the default parameters of the ComBat algorithm, and including age, sex, race, and time point in pregnancy in the model while using study accession as a surrogate for batch. Because SDY36 measured a smaller number of cytokines and cell subsets than are available as part of the 10KIP, we further selected a subset of the 10KIP to include just those parameters measured in SDY36. These data were used to conduct standard PCA analysis (R: prcomp, ggbiplot). Differences were calculated using ANOVA with a Tukey's HSD post hoc test.

Gene expression array harmonization and normalization—Gene expression array data were obtained in three formats. For data collected on Affymetrix platforms, we utilized the *ReadAffy* utility in the *affy* Bioconductor package to read in raw CEL files. The *rma* utility was used to conduct Robust Multichip Average (rma) background correction (as in (Irizarry etal., 2003)), quantile normalization, and log2 normalization of the data. For data collected on Illumina platforms and stored in the Gene Expression Omnibus (GEO) database, we utilized the *getGEO* utility in the *GEOquery* Bioconductor package to read the expression files and the *preprocessCore* package to conduction rma background correction, quantile normalization, and log2 normalization of the gene expression data. Finally, for data collected on Illumina platforms but not stored in GEO, we utilized the *read.ilmn* utility of the *limma* Bioconductor package to read in the data, and the *neqc* function to rma background correct, quantile normalize, and log2 normalize the gene expression data. In all instances, probe IDs were converted to Entrez Gene IDs. Where multiple probes mapped to the same Entrez Gene ID, the median value across probes was used to represent the expression value of the corresponding gene. The background-corrected and normalized

datasets were combined based on common Entrez IDs, missing values were imputed with a k-nearest neighbors algorithm (R package: impute, function: impute.knn) using k= 10 and default values for rowmax, colmax, and maxp. To create the normalized and batch corrected dataset available through the www.10kImmunomes.org portal, we utilized a well-established empirical Bayes algorithm for batch correction (Johnson et al., 2007), compensating for possible batch effects while maintaining potential effects of age, race, and sex across datasets and mapped Entrez IDs to HUGO gene IDs.

Simulations to validate the batch correction algorithm—The empirical Bayes algorithm we have used to generate the normalized data available for download has previously been validated in its use for gene expression microarray analysis (Johnson etal., 2007). To assess the efficacy of using an empirical Bayes algorithm to compensate for batch effects in multiplex ELISA data, we generated simulated multiplex ELISA data as skewed normal distributions from a set of parameters selected to mimic those skewed normal distributions that best fit the actual multiplex ELISA data used in our analysis. We generated this data for 50 analytes and 1500 subjects and purposefully introduced batch effects intended to mimic the types of batch effects we might encounter in real multiplex ELISA data. To account for use of a differently calibrated machine, for example, we simulated data in which one batch had a higher mean than the other batches. To account for the possibility that one lab's data might be more variable than others, in one simulation we introduced random noise into one batch of the data. Finally, to account for the fact that the antibodies used may differ in efficacy across lots and experiments, we devised a simulation in which just one analyte in just one batch has a perturbed mean. In each of 1000 simulations of this data, we then generated a linear model to test whether the empirical Bayes algorithm ComBat (Johnson et al., 2007) would successfully correct for these deviations from the true value of the simulated data. Additionally, we took the largest single batch of multiplex ELISA data (data from ImmPort study SDY 420) and intentionally introduced the same 3 types of batch effects we introduced into the simulated data. Following the same procedure, we demonstrate that ComBat successfully removes these introduced batch effects from real multiplex ELISA data.

QUANTIFICATION AND STATISTICAL ANALYSIS

Detailed descriptions of data collection and modeling are available in the Method Details section. All publicly available software is enumerated in the Key Resources table. Statistical tests used, p values of those tests, and the n of each test are detailed in the figure legends. In all cases, n represents the number of distinct human subjects represented in the test. In PCA and tSNE plots, each point represents an individual human subject. In violin plots, each black dash represents an individual human subject. The width of the violin can be read as a histogram, representing the density of subjects at each value. The length of the violin represents the range of values. In ribbon plots, each point represents an individual human subject, while ribbons indicate the mean and standard error of each group as a loess-smoothed curve, as implemented in the geom_smooth aesthetic of the R ggplot2 package. In forest plots, each point represents the effect size the variable of interest (age, sex, or race) has on the measured values, while the error bars represent the standard error of that estimate. In boxplots, the central line indicates the mean, the edges of the box represent the 25th and

75th percentile of the data. The upper whisker represents the smaller of the maximum value and the 75th percentile + 1.5 × the interquartile range. If the latter, outliers are represented by individual points. The lower whisker represents the larger of the minimum value and the 25th percentile −1.5 × the interquartile range. If the latter, outliers are represented by individual points.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank M. Sirota, H. Maecker, Y. Rosenberg-Hasson, M. Spitzer, and J. Puck for their guidance and early advice in this work. We thank B. Oskotsky for his key efforts in getting the 10kimmunomes.org website up and running. We would also like to acknowledge all of the original data contributors who, by submitting their individual-level data to ImmPort, made this work possible. Thisworkwas supported bythe National Institute ofAllergy and Infectious Diseases (Bioinformatics Support Contract HHSN272201200028C). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

REFERENCES

1000 Genomes Project Consortium; Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, and McVean GA (2010). A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073. [PubMed: 20981092]

1000 Genomes Project Consortium; Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, and McVean GA (2012). An integrated map ofgeneticvariation from 1,092 human genomes. Nature 491, 56–65. [PubMed: 23128226]

Ashley EA (2015). The precision medicine initiative: a new national effort. JAMA 313, 2119–2120. [PubMed: 25928209]

Becher B, Spath S, and Goverman J (2017). Cytokine networks in neuroinflammation. Nat. Rev. Immunol 17, 49–59. [PubMed: 27916979]

Belkaid Y, and Rouse BT (2005). Natural regulatoryTcells in infectious disease. Nat. Immunol 6, 353–360. [PubMed: 15785761]

Bolstad B (2017). preprocessCore: a collection of pre-processing functions. R package version 1.42.0. https://github.com/bmbolstad/preprocessCore.

Bhattacharya S, Andorf S, Gomes L, Dunn P, Schaefer H, Pontius J, Berger P, Desborough V, Smith T, Campbell J, et al. (2014). ImmPort: disseminating data to the public for the future of immunology. Immunol. Res 58, 234–239. [PubMed: 24791905]

Bianchi DW, Zickwolf GK, Weil GJ, Sylvester S, and DeMaria MA (1996). Male fetal progenitor cells persist in maternal blood for as long as 27 years postpartum. Proc. Natl. Acad. Sci. USA 93, 705–708. [PubMed: 8570620]

Brodin P, and Davis MM (2017). Human immune system variation. Nat. Rev. Immunol 17, 21–29. [PubMed: 27916977]

Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP, McCarthy MI, Ouwehand WH, Samani NJ, et al.; Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661–678. [PubMed: 17554300]

Chan WFN, Gurnot C, Montine TJ, Sonnen JA, Guthrie KA, and Nelson JL (2012). Male microchimerism in the human female brain. PLoS ONE 7, e45592. [PubMed: 23049819]

Chavey C, Lazennec G, Lagarrigue S, Clape C, Iankova I, Teyssier J, Annicotte J-S, Schmidt J, Mataki C, Yamamoto H, et al. (2009). CXC ligand 5 is an adipose-tissue derived factor that links obesity to insulin resistance. Cell Metab. 9, 339–349. [PubMed: 19356715]

Collins FS, and Varmus H (2015). A new initiative on precision medicine. N. Engl. J. Med 372, 793–795. [PubMed: 25635347]

Davis S, and Meltzer PS (2007). GEOquery: a bridge between the Gene Expression Omnibus(GEO) and BioConductor. Bioinformatics 23,1846–1847. [PubMed: 17496320]

Djoba Siawaya JF, Roberts T, Babb C, Black G, Golakai HJ, Stanley K, Bapela NB, Hoal E, Parida S, van Helden P, and Walzl G (2008). An evaluation of commercial fluorescent bead-based luminex cytokine assays. PLoS ONE 3, e2535. [PubMed: 18596971]

Dubuc GR, Phinney SD, Stern JS, and Havel PJ (1998). Changes of serum leptin and endocrine and metabolic parameters after 7 days of energy restriction in men and women. Metabolism 47, 429–434. [PubMed: 9550541]

Dunn PJ, Thomson E, Campbell J, Smith T, Desborough V, Wiser J, Schaefer H, Bhattacharya S, Butte AJ, Andorf S, et al. (2015). ImmPort: shared research data for bioinformatics and immunology. In 2015 IEEE International Conference on Bioinformaticsand Biomedicine (BIBM), pp. 607–610.

Ekouevi DK, Inwoley A, Tonwe-Gold B, Danel C, Becquet R, Viho I, Rouet F, Dabis F, Anglaret X, and Leroy V (2007). Variation of CD4 count and percentage during pregnancy and after delivery: implications for HAART initiation in resource-limited settings. AIDS Res. Hum. Retroviruses 23, 1469–1474. [PubMed: 18160003]

Elshal MF, and McCoy JP (2006). Multiplex bead array assays: performance evaluation and comparison of sensitivity to ELISA. Methods 38, 317–323. [PubMed: 16481199]

Finak G, Langweiler M, Jaimes M, Malek M, Taghiyar J, Korin Y, Raddassi K, Devine L, Obermoser G, Pekalski ML, et al. (2016). Standardizing flow cytometry immunophenotyping analysis from the Human Immuno-Phenotyping Consortium. Sci. Rep 6, 20686. [PubMed: 26861911]

Friedman AA, Letai A, Fisher DE, and Flaherty KT (2015). Precision medicine for cancer with next-generation functional diagnostics. Nat. Rev. Cancer 15, 747–756. [PubMed: 26536825]

Gautier L, Cope L, Bolstad BM, and Irizarry RA (2004). affy-analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20, 307–315. [PubMed: 14960456]

Hastie T, Tibshirani R, Balasubraman N, and Chu G (2017). Impute: imputation for microarray data. R package version 1.52.0 (Bioconductor)

Hu Z, Jujjavarapu C, Hughey JJ, Andorf S, Gherardini PF, Spitzer MH, Dunn P, Thomas CG, Campbell J, Wiser J, et al. (2017). Meta-analysis of cytometry data reveals racial differences in immune cells. bioRxiv. https://doi.org/10.1101/130948.

Huber JP, and Farrar JD (2011). Regulation ofeffector and memoryT-cell functions by type I interferon. Immunology 132, 466–474. [PubMed: 21320124]

Hughey JJ, and Butte AJ (2015). Robust meta-analysis ofgene expression using the elastic net. Nucleic Acids Res. 43, e79. [PubMed: 25829177]

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, and Speed TP (2003). Exploration, normalization, and summaries of high densityoligonucleotide array probe level data. Biostatistics 4,249–264. [PubMed: 12925520]

Johnson WE, Li C, and Rabinovic A (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8, 118–127. [PubMed: 16632515]

Joseph CL, Ownby DR, Peterson EL, and Johnson CC (2000). Racial differences in physiologic parameters related to asthma among middle-class children. Chest 117, 1336–1344. [PubMed: 10807820]

Khan SS, Smith MS, Reda D, Suffredini AF, and McCoy JP, Jr. (2004). Multiplex bead array assaysfordetection ofsoluble cytokines: comparisonsof sensitivity and quantitative values among kits from multiple manufacturers. Cytometry B Clin. Cytom 61, 35–39. [PubMed: 15351980]

Kirkpatrick DP, and Dransfield MT (2009). Racial and sex differences in chronic obstructive pulmonary disease susceptibility, diagnosis, and treatment. Curr. Opin. Pulm. Med 15, 100–104. [PubMed: 19532023]

Kraus TA, Engel SM, Sperling RS, Kellerman L, Lo Y, Wallenstein S, Escríbese MM, Garrido JL, Singh T, Loubeau M, and Moran TM (2012). Characterizing the pregnancy immune phenotype: results of the viral immunity and pregnancy (VIP) study. J. Clin. Immunol 32, 300–311. [PubMed: 22198680]

La Cava A, and Matarese G (2004). The weight of leptin in immunity. Nat. Rev. Immunol 4, 371–379. [PubMed: 15122202]

Lee EY, Lee Z-H, and Song YW (2009). CXCL10 and autoimmune diseases. Autoimmun. Rev 8, 379–383. [PubMed: 19105984]

Leek JT, and Storey JD (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 3, 1724–1735. [PubMed: 17907809]

Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, and Irizarry RA (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. Nat. Rev. Genet 11, 733–739. [PubMed: 20838408]

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291. [PubMed: 27535533]

Leng SX, McElhaney JE, Walston JD, Xie D, Fedarko NS, and Kuchel GA (2008). ELISA and multiplex technologies for cytokine measurement in inflammation and aging research. J. Gerontol. A Biol. Sci. Med. Sci 63, 879–884. [PubMed: 18772478]

Mackay CR (2001). Chemokines: immunology's high impactfactors. Nat. Immunol 2, 95–101. [PubMed: 11175800]

Maecker HT, Nolan GP, and Fathman CG (2010). New technologies for autoimmune disease monitoring. Curr. Opin. Endocrinol. Diabetes Obes 17, 322–328. [PubMed: 20531181]

Morgan E, Varro R, Sepulveda H, Ember JA, Apgar J, Wilson J, Lowe L, Chen R, Shivraj L,Agadir A, et al. (2004). Cytometric bead array: a multiplexed assay platform with applications in various areas of biology. Clin. Immunol 110, 252–266. [PubMed: 15047203]

Neurath MF (2014). Cytokines in inflammatory bowel disease. Nat. Rev. Immunol 14, 329–342. [PubMed: 24751956]

O'Shea JJ, Ma A, and Lipsky P (2002). Cytokines and autoimmunity. Nat. Rev. Immunol 2, 37–45. [PubMed: 11905836]

Ooms J, James D, DebRoy S, Wickham H, and Horner J (2018). RMySQL: database interface and "MySQL" driver for R (RStudio).

Pelz CR, Kulesz-Martin M, Bagby G, and Sears RC (2008). Global rankinvariant set normalization (GRSN) to reduce systematic distortions in microarray data. BMC Bioinformatics 9, 520. [PubMed: 19055840]

Procaccini C, De Rosa V, Galgani M, Carbone F, La Rocca C, Formisano L, and Matarese G (2013). Role of adipokines signaling in the modulation of T cells function. Front. Immunol 4, 332. [PubMed: 24151494]

Qiu Y, Zhu J, Bandi V, Atmar RL, Hattotuwa K, Guntupalli KK, and Jeffery PK (2003). Biopsy neutrophilia, neutrophil chemokine and receptor gene expression in severe exacerbations of chronic obstructive pulmonary disease. Am. J. Respir. Crit. Care Med 168, 968–975. [PubMed: 12857718]

Qiu Y, Zhu J, Bandi V, Guntupalli KK, and Jeffery PK (2007). Bronchial mucosal inflammation and upregulation of CXC chemoattractants and receptors in severe exacerbations of asthma. Thorax 62, 475–482. [PubMed: 17234659]

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, and Smyth GK (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47. [PubMed: 25605792]

Rosenberg-Hasson Y, Hansmann L, Liedtke M, Herschmann I, and Maecker HT (2014). Effects of serum and plasma matrices on multiplex immunoassays. Immunol. Res 58, 224–233. [PubMed: 24522699]

Saeys Y, Van Gassen S, and Lambrecht BN (2016). Computational flow cytometry: helping to make sense of high-dimensional immunology data. Nat. Rev. Immunol 16, 449–462. [PubMed: 27320317]

Sauteraud R, Dashevskiy L, Finak G, and Gottardo R (2016). Immune-Space: enabling integrative modeling of human immunological data. J. Immunol 196 (Suppl 1), 124.65.

Shay T, and Kang J (2013). Immunological Genome Project and systems immunology. Trends Immunol. 34, 602–609. [PubMed: 23631936]

Sirak T, Chiadika S, Daka M, and Simon C (2008). Acute coronary syndrome in African Americans and Hispanic Americans In Acute Coronary Syndrome, Hong MK and Herzog E, eds. (Springer), pp. 229–245.

Skilton MR, Nakhla S, Sieveking DP, Caterson ID, and Celermajer DS (2005). Pathophysiological levels of the obesity related peptides resistin and ghrelin increase adhesion moleculeexpression on humanvascularendothelial cells. Clin. Exp. Pharmacol. Physiol 32, 839–844. [PubMed: 16173945]

Spanakis EK, and Golden SH (2013). Race/ethnic difference in diabetes and diabetic complications. Curr. Diab. Rep 13, 814–823. [PubMed: 24037313]

Spitzer MH, and Nolan GP (2016). Mass cytometry: single cells, manyfeatures. Cell 165, 780–791. [PubMed: 27153492]

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH, et al.; 1000 Genomes Project Consortium (2015). An integrated map of structural variation in 2,504 human genomes. Nature 526, 75–81. [PubMed: 26432246]

Verma S, Li S-H, Wang C-H, Fedak PWM, Li R-K, Weisel RD, and Mickle DAG (2003). Resistin promotes endothelial cell activation: further evidence of adipokine-endothelial interaction. Circulation 108, 736–740. [PubMed: 12874180]

Vignali DAA (2000). Multiplexed particle-based flow cytometric assays. J. Immunol. Methods 243, 243–255. [PubMed: 10986418]

Vignali DAA, Collison LW, and Workman CJ (2008). How regulatory T cells work. Nat. Rev. Immunol 8, 523–532. [PubMed: 18566595]

Winer DA, Winer S, Shen L, Wadia PP, Yantha J, Paltser G, Tsui H, Wu P, Davidson MG, Alonso MN, et al. (2011). B cells promote insulin resistance through modulation of T cells and production of pathogenic IgG antibodies. Nat. Med 17, 610–617. [PubMed: 21499269]

Wynn TA (2003). IL-13effectorfunctions. Annu. Rev. Immunol 21, 425–456. [PubMed: 12615888]
Zineh I, Beitelshees AL, Welder GJ, Hou W, Chegini N, Wu J, Cresci S, Province MA, and Spertus JA (2008). Epithelial neutrophil-activating peptide (ENA-78), acute coronary syndrome prognosis, and modulatory effect of statins. PLoS ONE 3, e3117. [PubMed: 18769620]

Highlights

- Subject-level immunology data from >10,000 healthy normal human subjects
- Curated data are available in raw or batch-corrected and normalized formats
- Interactive visualizations and downloads at 10kimmunomes.org

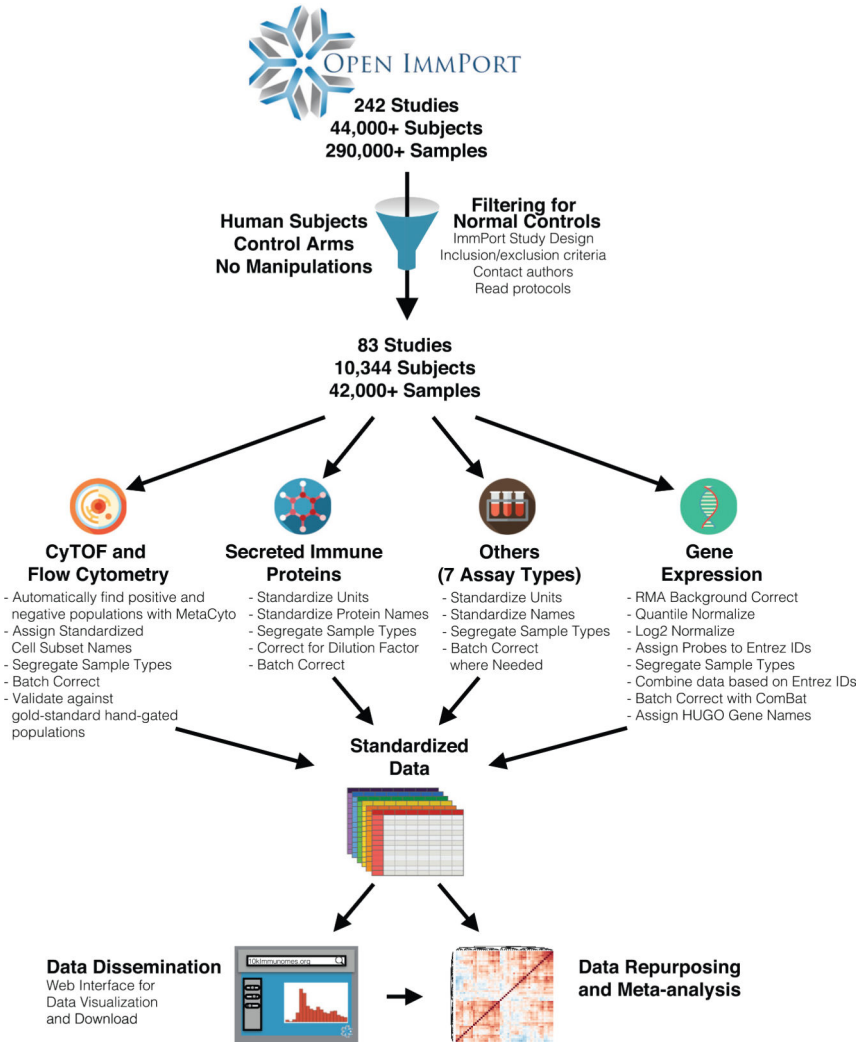


Figure 1. Resource Development and Selected Applications

Data from 242 studies and 44,775 subjects (including flow cytometry and CyTOF, mRNA expression, secreted protein levels [including cytokines, chemokines, and growth factors], clinical lab tests, HAI titers, HLA type, and others) were collected from the NIAID Immunology Data and Analysis Portal, ImmPort (http://www.immport.org/). We hand curated the entire contents of ImmPort to filter for normal healthy control human subjects. Each of the 10 data types was systematically processed and harmonized. These data constitute the largest compendium to date of cellular and molecular immune measurements on healthy normal human subjects. Both the normalized data and their raw counterparts are openly available for visualization and download at http://10kimmunomes.org/.

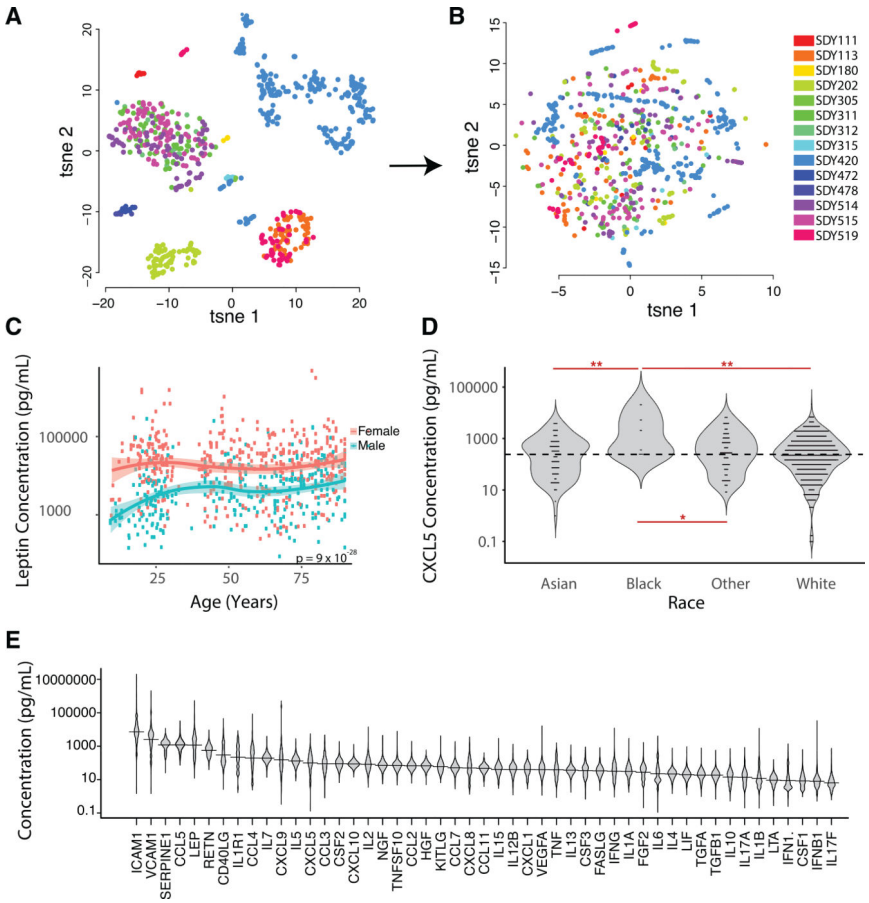


Figure 2.

High-Throughput Secreted Protein Data: Characterizing the Range of Unperturbed Secreted Protein Levels in a Diverse Population

(A) t-distributed stochastic neighbor embedding (tSNE) visualization of high-throughput secreted protein data, colored by study accession, reveals that much of the variance across the data is explained by batch.

(B) After batch correction with an empirical Bayes algorithm, which accounts for both mean and variance difference across studies while maintaining effects of covariates such as age, sex, and race, the data no longer cluster by batch.

(C) Secreted protein data as measured by multiplex ELISA across 17 studies captures known effects, such as elevated levels of serum leptin in female relative to male subjects (analysis of covariance [ANCOVA], n = 906, $p = 9 \times 10^{-28}$). Each point represents an individual subject. Ribbons indicatethe mean and standard error ofeach group.

(D) Analysis of the reference population reveals demographic associations, including elevated CXCL5 in African American subjects as compared to other races. (ANCOVA, n = 917, p values: *p < 0.05, **p < 0.01). Each dash represents an individual subject. The width of the violin represents the relative density of subjects at each value. The length of theviolin represents the range ofvalues.

(E) We characterize the distribution of secreted protein levels from serum across the reference population (n = 1,286). Each dash represents an individual subject. The width of

the violin represents the relative density of subjects at each value. The length of the violin represents the range of values.

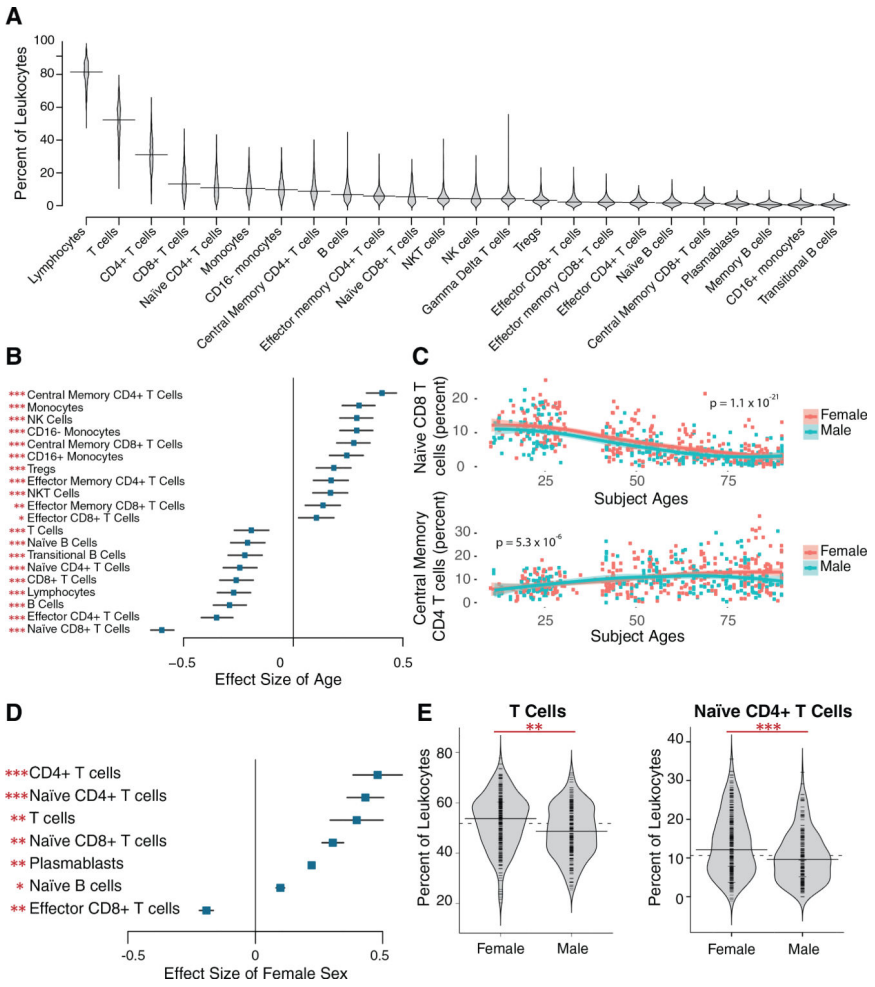


Figure 3.

Mass Cytometry: Characterizing the Range of Cell-Subset Percentages in a Diverse Population

(A) Distribution of cell-subset percentages across the 10KIP. The width of the violin represents the relative density of subjects at each value. The length of the violin represents the range of values.

(B) Analysis of mass cytometry data reveals significant effects of age on cell-subset percentages while accounting for sex and race. Only cell-subset associations with Benjamini-Hochberg-corrected p values < 0.05 are shown. (ANCOVA, n = 578, *p < 0.05, **p < 0.01, ***p < 0.001). Effect sizes are displayed as Pearson's r ± 95% confidence intervals. Each point represents the effect size of age. Error bars represent the standard error of that estimate.

(C) Naive CD8+ T cells decrease significantly with age (ANCOVA, n = 565, $p = 1.1 \times 10^{\sim 21}$), and central memory CD4 T cells increase significantly with age (ANCOVA, n = 578, $p = 5.3 \times 10^{-6}$), while accounting for sex and race. Each point represents an individual subject. Ribbons indicate the mean and standard error of each group.

(D) Analysis of mass cytometry data reveals significant effects of sex on cell-subset percentages, while accounting for age and race. Only cell-subset associations with

Benjamini-Hochberg-corrected p values < 0.05 are shown. (ANCOVA, n = 578, *p < 0.05, **p < 0.01, ***p < 0.001). Effect sizes are displayed as Cohen's d ± 95% confidence intervals. Each point represents the effect size age. Errorbars representthestandard errorofthat estimate.

(E) T cells (ANCOVA, n = 565, $p = 7.4 \times 10^{-6}$) and naive CD4+ T cells (ANCOVA, n = 578, $p = 3.3 \times 10^{-8}$) are significantly elevated in women as compared to men, accounting for age and race. Each dash represents an individual subject. The width of the violin represents the relative density of subjects at each value. The length of the violin represents the range of values.

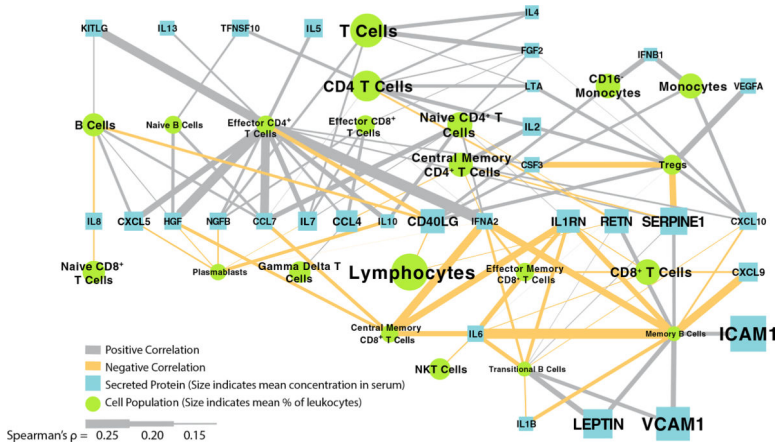


Figure 4. Immune Cell and Serum Cytokine Bipartite Graph

Immune cell percentages and serum protein concentrations, as measured by CyTOF and multiplex ELISA, were processed as described in **STAR Methods**, and the cell-cytokine relationship was described as partial correlations accounting for age, sex, and race. Only relationships significant at a BH-corrected $p < 0.01$ are shown.

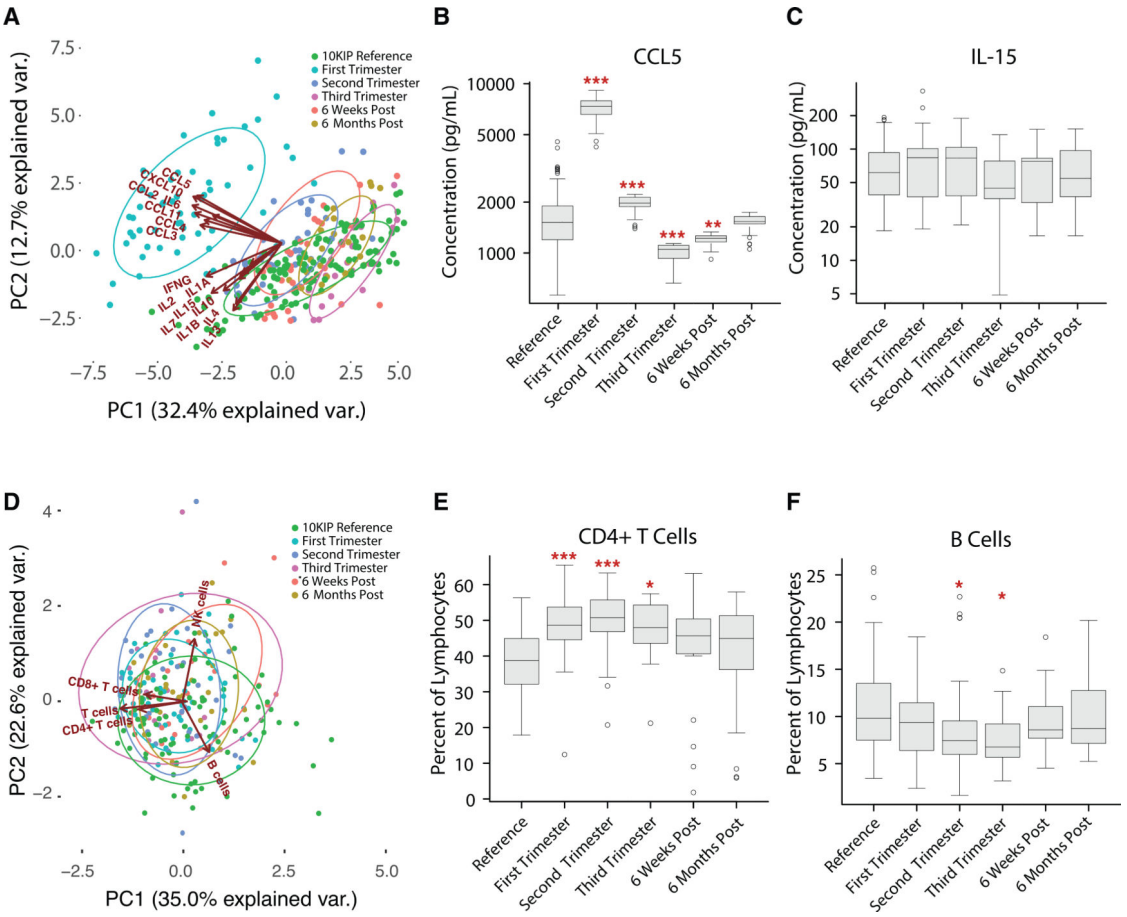


Figure 5. Comparing Pregnancy Data to the Common Control Reveals Cell-Subset and Immune Protein Modulation in Pregnancy

(A) PCA plot depicting the variation in serum proteins, as measured by multiplex ELISA, over the course of pregnancy, taken from ImmPort Study SDY36, as compared to multiplex ELISA measurements from women between the ages of18–40 from the reference population. The variance in measurements is dominated by a deviation in serum cytokine measurements during the first trimester(teal) relative to all other time points during pregnancy and relative to the 10KIP controls (green). These differences are driven primarily by changes in CCL2, CCL3, CCL4, CCL5, CCL11, IL6, and CXCL10.

(B) As an example of cytokine modulation in pregnancy, serum CCL5 levels are significantly increased in the first and second trimester relative to the 10KIP controls, decrease during the third trimester and remain low for at least 6 weeks postpartum. CCL5 levels return to baseline levels by6months postpartum (ANOVA with Tukey HSD, n = 142 controls, n = 57 pregnancy, *p < 0.05, **p < 0.01, ***p < 0.001).

(C) In contrast, serum IL15 levels make no significant deviations from normal over the course of pregnancy (ANOVA with Tukey HSD, n = 142 controls, n = 57 pregnancy).

(D) PCA plot depicting the variation in immune cell subsets, as measured by flow cytometry, over the course of pregnancy, taken from ImmPort Study SDY36, as compared to cytometry measurements from women between the ages of18and 40yearsfromthe 10KIPcontrols.Asopposed to cytokine measurements(A), the preponderance of variation in

cell-subset measurements is not due to changes over the course of pregnancy. All time points during and following gestation substantially overlap with the controls (green).

(E) The percentage ofCD4+ T cells, as a fraction of lymphocytes, is significantly elevated over the duration of pregnancy but returns to baseline in the postpartum period (ANOVA with Tukey HSD, n = 94 controls, n = 57 pregnancy, *p < 0.05, ***p < 0.001).

(F) The percentage of B cells, as a fraction of lymphocytes, exhibits a small but significant dip in the second and third trimesters (ANOVA with Tukey HSD, n = 94 controls, n = 57 pregnancy, *p < 0.05).

(B, C, E, and F) The central line indicates the mean. Edges of the box represent the 25th and 75th percentile of the data. The upper whisker represents the smaller of the maximum value and the 75th percentile + 1.5 × the interquartile range. If the latter, outliers are represented by individual points. The lower whisker represents the larger of the minimum value and the 25th percentile — 1.5 × the interquartile range. If the latter, outliers are represented by individual points

Table 1.

Data Available in the Initial Release

Measurement	Subjects
ELISA	4,035
Multiplex ELISA (Luminex)	1,286
Virus neutralization titer	2,265
HAI titer	1,344
Complete blood count	1,684
Comprehensive metabolic panel	664
Fasting lipid profile	664
Questionnaire	1,422
Flow Cytometry (PBMCs)	907
CyTOF (PBMCs)	583
Flow cytometry (whole blood)	164
HLA type	1,093
Gene expression (whole blood)	311
Gene expression (PBMCs)	165

Counts of distinct subjects for whom raw data of each type is represented in the initial release of the 10KIP. Because many subjects contributed multiple measurement types, the totals across all measurement types substantially exceed the number of distinct subjects. Data are available in the 10,000 Immunomes Project