Computational resources for high-dimensional immune analysis from the Human Immunology Project Consortium

To the Editor:
The human immune system forms a complex network of tissues, cells and molecules that protect against a wide variety of pathogens. It is sophisticated and adaptive in that it can distinguish self from non-self, and once it responds to a particular pathogen it exhibits 'memory', or heightened responsiveness, to that particular pathogen. This sophistication is facilitated in large part by the enormous diversity and large numbers of immune cell types, antigen receptors and cytokines. Although experiments in inbred mice housed in pathogen-free conditions have provided major insights into the mechanistic workings of the immune system, if we wish to understand how the immune system as a whole responds to a wide variety of pathogens, especially in outbred human populations, computational methods must be developed and databases created to track the many components and variables at play.

To meet this challenge, the National Institute of Allergy and Infectious Diseases (NIAID) of the US National Institutes of Health (NIH) created the Human Immunology Project Consortium (HIPC; http://www.immuneprofiling.org/). This competitive grants program currently consists of seven research centers, which are building large data sets on human subjects undergoing influenza vaccination or who are infected with pathogens including influenza virus, West Nile virus, herpes zoster, pneumococcus and the malaria parasite. Each HIPC research center also has biostatistics and bioinformatics experts who analyze and organize these data. These personnel also constitute a subcommittee that collaboratively works to create an infrastructure to assist the entire international immunology community; specifically this subcommittee has these goals:

- Development and implementation of standards for data collection, integration and data exchange
- Development of state-of-the-art algorithms and tools for the modeling and integration of heterogeneous immunological data
- Development and implementation of a central database and analysis engine providing easy access to all data generated through HIPC.

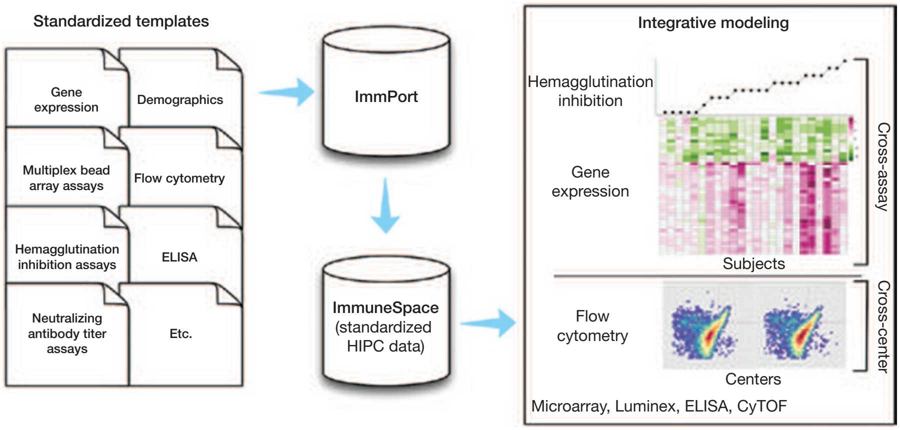
Standardized templates
Gene expression
Demographics
Multiplex bead array assays
Flow cytometry
Hemagglutination inhibition assays
ELISA
Neutralizing antibody titer assays
Etc.
ImmPort
ImmuneSpace (standardized HIPC data)
Integrative modeling
Hemagglutination inhibition
Gene expression
Subjects
Cross-assay
Flow cytometry
Centers
Cross-center
Microarray, Luminex, ELISA, CyTOF

Figure 1 HIPC data exchange workflow. Data are submitted to ImmPort by each HIPC center using a set of standardized data templates. Then, once public, these HIPC data are transferred to ImmuneSpace to enable integrative modeling across data types and HIPC centers.

A range of experimental protocols are available for measuring features of immune responses, including B- and T-cell specificity and repertoire, serum and intracellular cytokines, and signaling many of these immune parameters can be measured using ever-more sophisticated single-cell analysis techniques. Minimum information guidelines, which enable the unambiguous interpretation of the results of an experiment and facilitate the reproduction of the experiment, exist for general biological investigation[1], T-cell assays[2], microarray experiments[3] and flow cytometry[4]. Unfortunately, some immunological assays, such as multiplex bead array assays, lack data standards. Even where there are data standards (e.g., for flow cytometry), they are often not adopted by manufacturers and software companies.

To support the wide range of immunological experiments, HIPC is taking advantage of the considerable infrastructure already developed as part of the NIAID Immunology Database and Analysis Portal (ImmPort) system (https://immport.niaid.nih.gov/), which serves as a repository of data generated by investigators funded by the NIAID Division of Allergy, Immunology, and Transplantation. ImmPort facilitates data standardization because to submit data to

Published online 19 January 2014; doi:10.1038/nbt.2777.

ImmPort, experimental results and meta-data must be copied into templates to be uploaded to the system. HIPC is extending the existing set of ImmPort data submission templates by identifying explanatory information (known as "meta-data") that defines immunological experiments more completely. These include, for example, information on standard curves in Luminex experiments or experimental batches in gene expression microarray studies. Several of these HIPC data templates (including those that facilitate description of human subjects, biological samples and multiplex bead array assays) have now been adopted by ImmPort as improved standards.

One shortcoming of the current ImmPort system is that it is not currently compliant with biological ontologies[5], and therefore it is not yet structured as a directly computable form of knowledge. This means that each HIPC center could in theory use different terms to refer to the same thing, making cross-center data integration a serious challenge. For example, while collecting serum cytokine data for one cross-center project, we found that each of three centers involved referred to interleukin-2 by a different name (IL-2, IL2 and hIL-2). To address this problem, we associate data fields in the ImmPort templates with ontologies as a source of controlled vocabularies. For instance, cytokine names could be drawn from terms in the Protein Ontology. The use of controlled vocabularies ensures that data can be searched and integrated reliably across centers. Many of the terms in these vocabularies will be drawn from a few major ontologies: Gene Ontology (GO), Protein Ontology (PRO), Cell Ontology (CL) and Ontology for Biomedical Investigations (OBI)[5]. To implement these new data standards, HIPC is currently developing a Data Entry Mapping database that formally links data fields with ontological concepts and/or terms and supports the automated generation of "standards-aware" data templates. That said, although ImmPort and existing ontologies provide a basis for defining data standards, a substantial effort is needed to extend these standards across all types of HIPC-generated data.

Although the tools for collecting immunological data have advanced considerably in recent years, the tools for analyzing that data have not progressed nearly as much. In addition, we do not have a uniform and reliable way of integrating data sets across different assays (e.g., flow cytometry and gene expression), although a prototype of a data integration tool does exist[6]. With its inherent complexity and diversity, immunology presents unique bioinformatics and statistical challenges.

Several collaborative HIPC projects ranging from low-level analyses such as preprocessing and standardization of data, to high-level analyses that aim to integrate data across assays and/or HIPC centers are underway. For example, several projects focus on cytometry data. One project that aims to standardize flow cytometry data analysis is a collaborative effort between the FITMaN initiative[7] and the FlowCAP group[8]; this project seeks to ensure proper assay standardization from the data generation to the data analysis steps using automated gating[9]. Other projects focus on the more recently developed approach of mass cytometry[10,11] where 40 or more metal isotope labels can be used to interrogate single cells. HIPC centers have been very active in using this new technology and developing analytical methods[12] that are being integrated into our programs and database. Other specific analytical projects funded through HIPC include:

- Modeling bead-level data for advanced Luminex data analysis
- Deconvoluting high-bandwidth data to parse out cell-specific effects
- Creating an immune signatures collection for gene set enrichment analysis (http://www.broadinstitute.org/gsea/msigdb/collections.jsp#C7)
- Network modeling of vaccine responses
- Designing more advanced human leukocyte antigen typing methods[13,14]
- Developing ultrasensitive physical detection technologies to identify T-cell epitopes[15].

More details on these (completed or ongoing) projects are available on our consortium website.

A central HIPC goal is sharing data as widely and freely as possible to promote exploratory, descriptive and predictive research. To facilitate data sharing and analysis, HIPC has implemented a centralized database and analysis engine, ImmuneSpace (http://www.immunespace.org/), based on the LabKey system[16]. Data from ImmPort are automatically loaded into ImmuneSpace and joined with basic metadata (e.g., cohort membership, treatment information) to facilitate data exploration, visualization and analyses (**Fig. 1**). Although development is ongoing, ImmuneSpace already has a powerful programming interface that can be used to assemble complex data analysis pipelines using, for example, the R programming language[17]. ImmuneSpace also makes use of the Bioconductor project[18] to provide standardized analysis workflows written as dynamic reports using a customized interface to *knitr*[19], a popular package for reproducible report generation. This enables coding in R, with results and figures embedded in a web page within ImmuneSpace to provide complete transparency to users.

An example of a simple workflow is one that will take raw gene expression data from a vaccination experiment and produce normalized gene expression data with a list of differentially expressed genes at different time points. This workflow framework is very flexible and provides a mechanism for data validation, standardization and quality control to ensure accuracy, consistency, and compliance with standards, as well as full reproducibility of intermediate and final results. ImmuneSpace will also integrate many of the HIPC-generated analytical tools mentioned above.

In conclusion, by helping create necessary computational infrastructure and tools, HIPC has made substantial progress toward enabling a systems approach to immunological analysis. Although we are not the only group working on systems immunology—for example, the NIH intramural program supports the Center for Human Immunology, Autoimmunity and Inflammation (http://www.nhlbi.nih.gov/resources/chi/)—we are the only group that we know of working on generating a centralized public database and analysis engine for disseminating standardized human data and facilitating data exploration and analyses. We also would very much like to invite anyone else with these interests to join us in making this as useful a resource as possible. With additional effort from HIPC and the community, we feel confident that systems immunology will become a distinct and valuable specialty within immunology, much as genomics and the study of genomes has become an indispensable part of genetics.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper (doi:10.1038/nbt.2777).

***Vladimir Brusic*[1,7], *Raphael Gottardo*[2,7], *Steven H Kleinstein*[3,4], *Mark M Davis*[5] *& the HIPC steering committee*[6]**

[1]*Cancer Vaccine Center, Dana-Farber Cancer Institute, Boston, Massachusetts, USA.* [2]*Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA.* [3]*Interdepartmental Program in Computational Biology and*

Bioinformatics, Yale University, New Haven, Connecticut, USA. [4]Department of Pathology, Yale School of Medicine, New Haven, Connecticut, USA. [5]The Howard Hughes Medical Institute, The Institute for Immunity, Transplantation and Infection, and The Department of Microbiology and Immunology, Stanford University, Stanford, California, USA. [6]Full list of members and affiliations appears at the end of the paper. [7]These authors contributed equally to this work.
e-mail: rgottard@fhcrc.org and vladimir_brusic@dfci.harvard.edu

The members of the HIPC steering committee are as follows: Mark M Davis[5], David A Hafler[8], Helen Quill[9], A Karolina Palucka[10], Gregory A Poland[11], Bali Pulendran[12], Ellis L Reinherz[1], Kenneth D Stuart[13] & Alkis Togias[9]

[8]Departments of Neurology and Immunobiology, Yale University, New Haven, Connecticut, USA. [9]National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland, USA. [10]Baylor Institute for Immunology Research, Baylor Research Institute, Dallas, Texas, USA. [11]Mayo Vaccine Research Group, Division of General Internal Medicine, Mayo Clinic, Rochester, Minnesota, USA. [12]Department of Pathology and Laboratory Medicine, Emory University, Atlanta, Georgia, USA. [13]Seattle Biomed, Seattle, USA.

1. Taylor, C.F., Field, D., Sansone, S.A. & Aerts, J. *Nat. Biotechnol.* **26**, 889–896 (2008).
2. Britten, C.M. *et al. Immunity* **37**, 1–2 (2012).
3. Brazma, A. *et al. Nat. Genet.* **29**, 365–371 (2001).
4. Lee, J.A. *et al. Cytometry A* **73A**, 926–930 (2008).
5. Ceusters, W. & Smith, B. *Stud. Health Technol. Inform.* **160**, 1050–1054 (2010).
6. Siebert, J.C., Munsil, W., Rosenberg-Hasson, Y., Davis, M.M. & Maecker, H.T. *J. Transl. Med.* **10**, 62 (2012).
7. Maecker, H.T., McCoy, J.P. & Nussenblatt, R. *Nat. Rev. Immunol.* **12**, 191–200 (2012).
8. Aghaeepour, N. *et al. Nat. Methods* **10**, 228–238 (2013).
9. Lo, K., Brinkman, R.R. & Gottardo, R. *Cytometry A* **73A**, 321–332 (2008).
10. Newell, E.W., Sigal, N., Bendall, S.C., Nolan, G.P. & Davis, M.M. *Immunity* **36**, 142–152 (2012).
11. Horowitz, H. *et al. Sci. Transl. Med.* **5**:208ra145 (2013).
12. Amir, E.-A.D. *et al. Nat. Biotechnol.* **31**, 545–552 (2013).
13. Wang, C. *et al. Proc. Natl. Acad. Sci. USA* **109**, 8676–8681 (2012).
14. Lank, S.M. *et al. BMC Genomics* **13**, 378 (2012).
15. Reinhold, B., Keskin, D.B. & Reinherz, E.L. *Anal. Chem.* **82**, 9090–9099 (2010).
16. Nelson, E.K. *et al. BMC Bioinformatics* **12**, 71 (2011).
17. Ihaka, R. & Gentleman, R.R. *J. Comput. Graph. Stat.* **5**, 299 (1996).
18. Gentleman, R.C. *et al. Genome Biol.* **5**, R80 (2004).
19. Xie, Y. *Dynamic Report Generation with R and knitr* (Chapman & Hall, 2013).