© 2008 Nature Publishing Group http://www.nature.com/natureimmunology

The Immunological Genome Project: networks of gene expression in immune cells

Tracy S P Heng, Michio W Painter & The Immunological Genome Project Consortium

The Immunological Genome Project combines immunology and computational biology laboratories in an effort to establish a complete 'road map' of gene-expression and regulatory networks in all immune cells

The immune system is a 'cat's cradle' of networks operating at various levels, comprising a network of genetic and signaling pathways subtending a network of interacting cells. Hence, understanding the role of a given molecule or pathway in immune system function requires deciphering its effect in the context of these many networks. As two thirds of the genome is active in one or more immune cell type(s), with less than 1% of genes expressed exclusively in a given type of cell[1], the phenomena and molecules studied should be considered in the framework of the system as a whole. In addition, any given molecule can have opposite or paradoxical effects on functional outcome depending on its location or the other genes or gene products it interacts with. Fixating studies on classical immunity-related genes limits the understanding of a cell's function to what is already known at the risk of missing other genes that may function in immune responses. Narrowing the frame of cellular reference to one or a few characteristics or applying facile but misleading labels can lead to dangerously simplified paradigms. In this context, the discovery-driven approach of genomics is a key complement to hypothesis-driven experimentation. Conversely, immunology is an ideal field for the application of systems approaches, with its detailed descriptions of cell types (over 200 immune cell types are defined in the scope of the Immunological Genome Project (ImmGen)), wealth of reagents and easy access to cells.

Thanks to the broad and robust approaches allowed by gene-expression microarrays and related techniques, the transcriptome is probably the only '-ome' that can be reliably tackled in its entirety. Generating a complete perspective of gene expression in the immune system

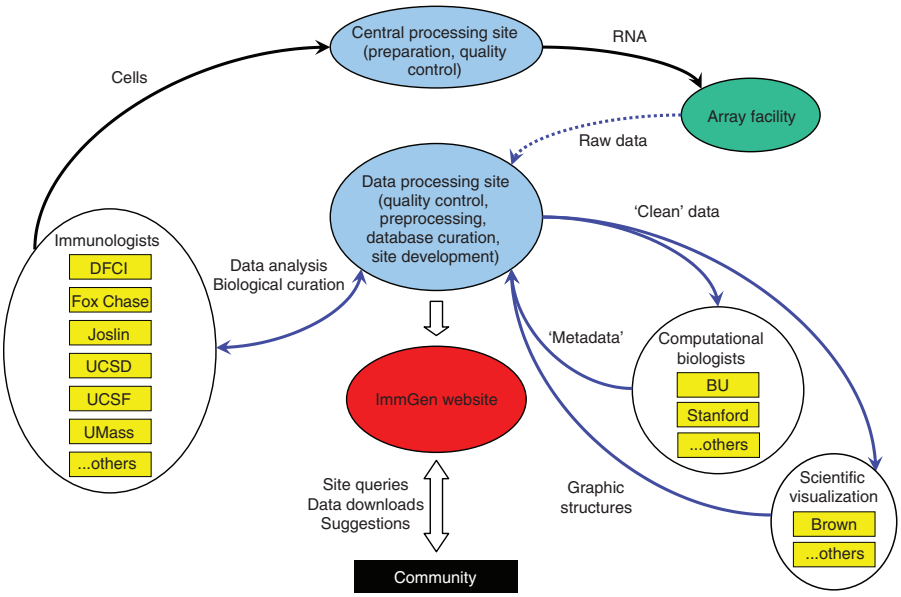

Figure 1 ImmGen consortium organization and project workflow. Cells purified by the immunology labs are sent to a central sample processing site for RNA extraction and checking and then are sent to a second site for amplification, labeling and hybridization. Quality-control tests are done on the raw data at a central data processing site, and confirmed data are then accessed by the immunology and computational biology labs for network analysis and development of 'visualization metaphors'. All data and metadata are accessible to the public through an Internet interface. DFCI, Dana-Farber Cancer Institute; Fox Chase, Fox Chase Cancer Center; Joslin, Joslin Diabetes Center; UCSD, University of California, San Diego; UCSF, University of California, San Francisco; UMass, University of Massachusetts; BU, Boston University; Stanford, Stanford University; Brown, Brown University,

Tracy S.P. Heng and Michio W. Painter are in the Section on Immunology and Immunogenetics, Joslin Diabetes Center & Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02215, USA.
A complete list of authors and affiliations appears at the end of this paper.
e-mail: immgen@joslin.harvard.edu

© 2008 Nature Publishing Group http://www.nature.com/natureimmunology

offers the potential for deciphering patterns that mirror responses at several levels. At the level of the gene, such data can give insight into how individual genes act along differentiation profiles and cellular responses. It is then possible to define modules or groups of genes whose expression is interdependent and is coordinated by shared regulatory controls. Computational methods for reverse engineering can then be applied to infer a model of the cells' underlying control system. Finally, genome-wide expression data at the highest level of integration provides an objective definition of the relations and distinctions between cells. For example, analyses of relative 'distances' in genomic space have shown that natural killer T cells are actually a subset of conventional $CD4^+$ T cells and not an 'intermediate' between $CD4^+$ T cells and natural killer cells, as is often believed[1]. Thus, insights from genomic profiling may 'fine tune' or revise the classifications and mental representations of immune cells.

Such goals require a coordinated effort on a large scale beyond the scope of any single laboratory. Many focused microarray studies have been done in the context of immunology, addressing the development and differentiation of various immunological lineages[2–4], characteristics of functional states[5,6] and perturbations associated with autoimmunity, immunopathology or malignancies[7–9]. However, there is a paucity of studies addressing gene-expression data across a substantial range of lineages. Microarray explorations can be robust but are very sensitive to experimental 'noise'[10], and the high degree of variation between platforms or laboratories undermines any direct comparison between data sets deposited into data warehouses such as the National Institutes of Health's GEO database or the European Molecular Biology Laboratory's ArrayExpress database. Some bioinformatic techniques have been proposed to overcome such variations in pairwise comparisons, but the remaining 'noise' renders any large-scale integration dubious[11]. Some compendia do exist, such as Symatlas[12], Immune Response In Silico[13], Genopolis[14] and the Reference Database of Immune Cells[15], but these sources are either too broadly or too narrowly focused, are incomplete or may not have sufficiently robust data to allow a comprehensive analysis of the immune system.

The broader goal of estimating a global regulatory network in a mammalian genome requires vast quantities of data with discrete perturbations that help unmask fine regulatory effects that would otherwise be hidden in data sets focusing on certain immune cell types[16,17]. Such analyses in the much smaller genome of yeast have required over 500 microarrays, and we estimate that over 2,000 data sets will be needed for analysis of the mouse genome. In addition, the quality of the data sets must be carefully controlled so that the biological signal is not overshadowed by 'noise' from lab and/or batch variability. Thus, a comprehensive genomic perspective of the immune system is not yet available. This is what the ImmGen group aims to address, robustly and comprehensively.

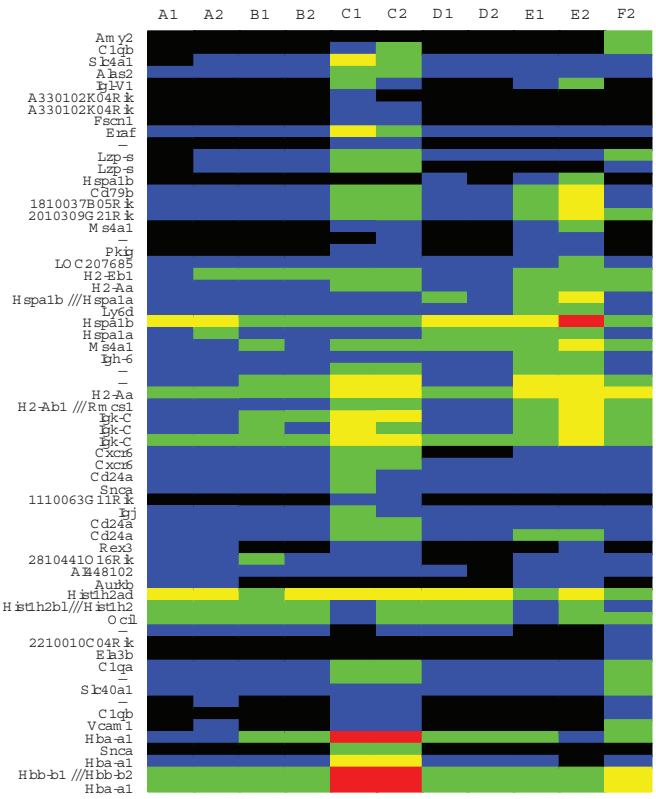

Figure 2 Heat-map representation of 'problem transcripts' from replicate spleen $CD4^+$ T cell samples (1 or 2) purified by each laboratory of the consortium (A–F) during the validation phase of the ImmGen profiling project. Results represent expression differences among data sets, presented on a spectrum ranging from low variability (black) to high variability (red). Some transcripts correspond to contamination by B cells (IgK, Igh-6 and H2-A[a]), erythrocytes (Hba and Hbb) or cells of unknown origin (Amy2) or indicate stress (Hsp).

Overall goals and organization of ImmGen

This project will generate, with rigorously standardized conditions, a complete compendium of genome-wide data sets showing the expression of protein-coding genes for all defined cell populations of the mouse immune system. The project will focus on primary cells isolated *ex vivo* in steady-state conditions or in response to genetic or environmental perturbations. Integrative computational tools will be applied to the expression profiles to reverse-engineer or predict the regulatory network in immune cells. Variation will be introduced into the analysis, through natural genetic polymorphism, knockout of genes, knockdown by RNA-mediated interference, or drug treatment, to drive and refine the computational network construction. For practical considerations, only the mouse genome will be studied at present, although future research may include human samples. All data and metadata (such as modules, 'signatures' and networks) will be made accessible to the public and the project will explore new visualization tools to support the display and browsing of genes, modules and connectivity.

The core group of the project comprises seven immunology and three computational biology laboratories (**Fig. 1**). Over 200 cell populations have been parsed by the immunology labs, each in charge of identifying and purifying all populations and subpopulations of its own cell grouping (under the umbrella of building the general compendium, each lab is also seeking to answer a series of focused questions related to their cells of interest). To minimize 'noise' from lab-specific variation, contamination, circadian effects, cell stress and so on, a common and strictly defined standard operating protocol is followed by all, with mice that originate from the same source (the Jackson Laboratory). For homogeneity, RNA preparation, probe labeling and hybridization are done in a centralized way. Quality checks and robust normalization are done on the raw data, which are then used by the computational biologists to analyze genetic modules and 'signatures' and to reverse-engineer underlying gene-regulatory networks, thus identifying genetic mediators for various immunological processes. In addition, ImmGen has ties with 'systems immunology' efforts in Europe and Asia. A collaboration with the European Union–supported Systems Biology on T-cell Activation consortium is focused on the fine-grained analysis of events occurring at various stages of T cell activation, and discussions are

© 2008 Nature Publishing Group http://www.nature.com/natureimmunology

underway with the Japanese RIKEN institutes for complementary analyses of the immunological transcriptome.

ImmGen is intended to be an open project. Beyond following guidelines of the National Human Genome Research Institute to allow rapid public access to the data and metadata, the group welcomes suggestions from the community about additional populations to profile (including direct participation in the form of reagents or help with cell preparations), community participation or suggestions for data analysis or the development of the web interface. The project is supported mainly by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health, but we have been fortunate in obtaining considerable support from suppliers of materials, mainly eBiosciences and Affymetrix (as of this writing).

Technical platforms

At present, only microarray or related hybridization-based technologies, on beads or a solid support, offer the breadth, throughput and reliability required for genome-wide profiling of protein-coding genes. Other techniques can be used for expression analysis, but these are of limited breadth (RT-PCR), are not cost-effective on a high-throughput scale (serial analysis of gene expression) or have yet to be proven reliable for such a study (high-throughput serial analysis of gene expression; color-coded single-molecule imaging). To provide uniformity across data sets and allow investigators to reference further data against ImmGen, the primary microarray platform will be held constant for the duration of the project and beyond. This primary platform for the analysis of protein-coding genes will be complemented by several additional investigations, which will be done on a more restricted set of 12 main leukocyte populations. Any microarray platform may fail to detect a certain amount of transcripts known to be present (false-negative results); this can be eliminated by analysis on a second array platform or new technologies that might arise during the course of the project. New-generation arrays that distinguish splicing variants will be used to explore alternative splicing across immunocytes. Transcription start sites will be analyzed (in collaboration with RIKEN groups). Once the cost and ease of the techniques have matured, massive parallel sequencing approaches for 'transcript counting' will be applied to cross-confirm or refine the microarray results for expression and alternative splicing. Although the main focus is on protein-coding genes, the importance of noncoding RNAs (microRNAs and others) for immune function is now well appreciated[18]. The knowledge of the genomic diversity and the reliability of analysis techniques are not as mature and robust for noncoding RNAs as they are for protein-coding genes, but ImmGen will profile them once the dust has settled.

Visualization of genome-scale data

Perhaps one of the greatest challenges in systems-level studies lies in organizing and visualizing complex metadata. This difficulty arises not simply from trying to represent large amounts of data in a user-friendly way but also in trying to determine and visually prioritize what data are meaningful in the broader context of a system. ImmGen has created and is developing new interactive tools to make data visually 'graspable' for genes and for more complex structures such as coregulated modules or 'signatures'. The website will also support outside queries, such as determining which cell type a particular profile most resembles, or which module or 'signature' distinguishes certain data sets. Once a complete immune gene-expression network has been established, individual investigators may query the network to specifically alter the activity of a given gene and catalog the ripple effects. Hence, the database will be an evolving entity that enables and calls for community participation.

ImmGen profiling: present status

The project is now operational. Its overall success is dictated by the quality and reproducibility of cell purifications, and initial studies have served to confirm and refine procedures in the participating labs. In early confirmation studies, profiles from the same cell population collected by all participating labs showed a large amount of interlab variability, with a smaller 'distance' for intralab than for interlab replicates (**Table 1**). This was unexpected, as the leaders of the laboratories have collectively 95 years of experience with cell sorting, and the target population being purified (spleen $CD4^+$ T cells) was not thought to present a challenge. The profiles showed 'signatures' of some expected sources of error (cell stress and contamination by B cells or erythrocytes) but also of unexplainable lab-specific transcripts (**Fig. 2**). These differences were minimized by further standardization of the sorting protocols, but they served to show that reliable and reproducible microarray data are achievable only with high sorting purity and strict adherence to a fixed protocol.

To select the microarray platform, we compared the sensitivity, 'noise', differential expression and reliability of detection of four commercial arrays using common RNA pools from $CD4^+$ and $CD19^+$ cells. The analyses focused on a subset of probes representing 12,297 genes common to all four arrays (matching GeneSymbol or the National Center for Biotechnology Information GeneID). We found distinct differences between arrays in terms of sensitivity (most platforms had 3–5% false-negative results; one had up to 15%), 'noise' (inter-replicate coefficients of varation ranging from 0.09 to 0.22) and the ability to detect differential expression (unexpectedly, FoldChange metrics proved globally different). Although no single array was the winner in all categories, we chose the Affymetrix Gene ST 1.0 array as the primary platform for the project. With the validation tests completed, the data generation for the first compendium phase is now underway. As of July 2008, the group had generated the first 150 data sets for 50 cell populations.

Conclusion

The genome-sequencing and variation-mapping projects have established essential genetic 'road maps'. In the same vein, if on a more focused scale, by generating robust expression data and metadata into a centralized and accessible location, ImmGen should provide an essential 'workbench' for delineating the intricate workings of the immunological genome.

Table 1 Genes with expression change between and within labs

	A1	A2	B1	B2	C1	C2	D1	D2	E1	E2	F1
A1	0	8	19	15	13	5	41	66	59	23	41
A2	4	0	20	17	21	3	22	33	94	8	54
B1	33	110	0	16	61	137	230	410	207	131	21
B2	14	10	3	0	18	9	38	91	92	15	19
C1	99	109	126	97	0	32	135	161	99	74	82
C2	86	81	101	90	12	0	107	128	111	56	72
D1	39	14	74	61	110	28	0	18	44	85	99
D2	6	8	15	8	24	3	6	0	42	22	42
E1	101	102	86	129	92	46	101	144	0	28	109
E2	95	111	117	113	67	59	159	174	68	0	128
F1	57	116	45	84	25	26	156	264	103	58	0

Genes with a change in expression of over twofold for intralab replicates (1 or 2) or interlab samples (A–F).

ACKNOWLEDGMENTS

We thank eBiosciences, Affymetrix and Expression Analysis for participation and support. Supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health (R24 AI072073 to C.B.).

1. Hyatt, G. *et al. Nat. Immunol.* **7**, 686–691 (2006).
2. Hoffmann, R., Bruno, L., Seidl, T., Rolink, A. & Melchers, F. *J. Immunol.* **170**, 1339–1353 (2003).
3. Mick, V.E., Starr, T.K., McCaughtry, T.M., McNeil, L.K. & Hogquist, K.A. *J. Immunol.* **173**, 5434–5444 (2004).
4. Edwards, A.D. *et al. J. Immunol.* **171**, 47–60 (2003).
5. Fontenot, J.D. *et al. Immunity* **22**, 329–341 (2005).
6. Kaech, S.M., Hemby, S., Kersh, E. & Ahmed, R. *Cell* **111**, 837–851 (2002).
7. Matos, M., Park, R., Mathis, D. & Benoist, C. *Diabetes* **53**, 2310–2321 (2004).
8. Bennett, L. *et al. J. Exp. Med.* **197**, 711–723 (2003).
9. Ebert, B.L. & Golub, T.R. *Blood* **104**, 923–932 (2004).
10. Shi, L. *et al. Nat. Biotechnol.* **24**, 1151–1161 (2006).
11. Bammler, T. *et al. Nat. Methods* **2**, 351–356 (2005).
12. Su, A.I. *et al. Proc. Natl. Acad. Sci. USA* **101**, 6062–6067 (2004).
13. Abbas, A.R. *et al. Genes Immun.* **6**, 319–331 (2005).
14. Splendiani, A. *et al. BMC Bioinformatics* **8** 31, S21 (2007).
15. Hijikata, A. *et al. Bioinformatics* **23**, 2934–2941 (2007).
16. di Bernardo, D. *et al. Nat. Biotechnol.* **23**, 377–383 (2005).
17. Segal, E. *et al. Nat. Genet.* **34**, 166–176 (2003).
18. Lodish, H.F., Zhou, B., Liu, G. & Chen, C.Z. *Nat. Rev. Immunol.* **8**, 120–130 (2008).

The complete list of authors is as follows (arranged by institution in random order; participants listed alphabetically by institution with principal investigator(s) listed last for each):

Kutlu Elpek[1], Veronika Lukacs-Kornek[1], Nora Mauermann[1], Shannon J Turley[1,11], Daphne Koller[2,11], Francis S Kim[3], Amy J Wagers[3,11], Natasha Asinovski[4], Scott Davis[4], Marlys Fassett[4], Markus Feuerer[4], Daniel H D Gray[4], Sokol Haxhinasto[4], Jonathan A Hill[4], Gordon Hyatt[4], Catherine Laplace[4], Kristen Leatherbee[4], Diane Mathis[4,11], Christophe Benoist[4,11], Radu Jianu[5], David H Laidlaw[5,11], J Adam Best[6], Jamie Knell[6], Ananda W Goldrath[6,11], Jessica Jarjoura[7], Joseph C Sun[7], Yanan Zhu[7], Lewis L Lanier[7,11], Ayla Ergun[8], Zheng Li[8], James J Collins[8,11], Susan A Shinton[9], Richard R Hardy[9,11], Randall Friedline[10], Katelyn Sylvia[10] & Joonsoo Kang[10,11]

[1]Department of Cancer Immunology and AIDS, Dana-Farber Cancer Institute, and Department of Pathology, Harvard Medical School, Boston, Massachusetts 02115, USA. [2]Department of Computer Science, Stanford University, Stanford, California 94305, USA. [3]Section on Developmental and Stem Cell Biology, Joslin Diabetes Center, Boston, Massachusetts 02115, USA, and Department of Stem Cell and Regenerative Biology, Harvard University and Harvard Stem Cell Institute, Cambridge, Massachusetts 02138, USA. [4]Section on Immunology and Immunogenetics, Joslin Diabetes Center, and Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. [5]Department of Computer Science, Brown University, Providence, Rhode Island 02912, USA. [6]Division of Biological Sciences, University of California, San Diego, La Jolla, California 92093, USA. [7]Department of Microbiology & Immunology and the Cancer Research Institute, University of California, San Francisco, San Francisco, California 94143, USA. [8]Department of Biomechanical Engineering and Center for BioDynamics, Boston University, Boston, Massachusetts 02215, USA. [9]Division of Basic Science, Fox Chase Cancer Center, Philadelphia, Pennsylvania 19111, USA. [10]Department of Pathology, University of Massachusetts Medical School, Worcester, Massachusetts 01655, USA. [11]Principal investigator.

© 2008 Nature Publishing Group http://www.nature.com/natureimmunology