HHS Public Access

Author manuscript

Trends Immunol. Author manuscript; available in PMC 2015 October 21.

Published in final edited form as:

Trends Immunol. 2013 December ; 34(12): 602–609. doi:10.1016/j.it.2013.03.004.

Immunological Genome project and systems immunology

Tal Shay[1] and **Joonsoo Kang**[2]

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA

[2]Department of Pathology, University of Massachusetts Medical School, 55 Lake Avenue North, Worcester, Massachusetts 01655, USA

Abstract

Immunological studies of single proteins in a single cell type are complemented in recent years by larger scale studies, enabled by emerging high-throughput technologies. This trend is recently exemplified by the discovery of gene networks controlling regulatory and effector αβ T cell subset development and human hematopoiesis. The Immunological Genome project aims to decipher the gene networks underpinning mouse hematopoiesis. The first phase, completed in 2012, profiled the transcriptome of 249 immune cell types. We discuss utilities of the datasets in high resolution mapping of the hematopoietic system. The immune transcriptome compendium has revealed unsuspected cell lineage relationships and the network reconstruction has identified novel regulatory factors of hematopoiesis.

A transcriptome compendium of mouse hematopoiesis

Classic immunology studies with a laser focus on a particular protein or biological process are becoming increasingly complemented by systems immunology studies that provide robust insights to fully understand the inner workings of the immune system. With technological advances, a systems immunology approach is feasible for individual laboratories, and not just for large consortia. However, the scope of individual enterprise still remains mostly restricted to a particular cell lineage [1-3], and as a consequence, variations on shared modular gene networks embedded in multiple cell components of a system are largely inaccessible. These variations are often the main drivers of diversity in cell types and the basis for division of labor within a cell system. As examples, cell type-specific function of SMAD3 downstream of TGFβ is dictated by lineage-specific master transcription factors (TFs) [4] and context-dependent functions of IRF4 can be accounted for by its association with distinct AP-1 factors [2, 5]. In practice, the ability to survey widely and yet recognize patterns of biological significance as they emerge necessitates a collective, in which uniform data generation and quality control go hand-in-hand with input from immunologists and computational biologists who can process the data and articulate clear paths towards deducing biological meanings of the emerging patterns[6]. There are several

Corresponding author: Kang, J. (Joonsoo.Kang@umassmed.edu).

initiatives that try to achieve this goal. For example, RefDIC, an open resource compendium of quantitative mRNA and Protein profile data of immune cells[7], and the Immunological Genome project (ImmGen), which is arguably the most comprehensive effort to date, and will be the focus of this review.

ImmGen is a consortium whose primary objective is to establish a comprehensive, public compendium of gene networks operating in the mouse hematopoietic system- the components, their properties, and their behaviors upon perturbations [8]. The consortium is made up of immunologists with expertise in specific immune cell subsets working with computational biologists in pursuit of network models to determine gene circuit architectures underpinning complex biological systems. Data so far has been restricted to RNA as it is the only reproducibly quantifiable trait of cell subsets present at limiting numbers in a system wide scale. Affymetrix microarrays were chosen as the data collection platform [8]. Alternative assays, such as RNA-seq, have been used to verify the data quality (http://www.immgen.org/Protocols), and upon optimization are likely to be more prominently utilized in the future. The ImmGen site (www.immgen.org) is a portal with several data browsers, each allowing the user to study a unique aspect of the ImmGen dataset (Table 1, Text box 1). There were 78,038 visits to the ImmGen site from 98 countries during 2012.

While the overarching thrust of ImmGen has been to map at high resolution the global interconnectivity of gene networks across cell constituents, solutions to investigator-driven, cell type-specific queries have fortified the network tree with rich details. Here we highlight newly emerging themes in the gene circuitry driving the hematopoietic system, constructed from the ImmGen mRNA profiles data amassed over the last six years, in large part using the network modeling program termed Ontogenet specifically developed for ImmGen (Fig. 1a). We first describe Ontogenet and the regulatory program it predicts, and then present examples for each of the three major thematic studies of ImmGen to date. We also discuss the integration of ImmGen with other datasets.

Ontogenet - a novel method for reconstructing regulatory networks in tree structured datasets

Ontogenet is a new method that combines linear regression with the tree structure of the dataset to predict a set of transcriptional regulators that would best account for each module's expression[9]. A module is a set of genes that are co-expressed across the dataset. Regulators are selected from a predefined list of factors that regulate gene transcription (TFs and chromatin modifiers). Ontogenet is specifically devised to address some of the challenges – and leverage some of the unique power – of studying transcriptional programs in cell lineages. First, Ontogenet can identify a whole set of ‘equivalent’ regulators, whereas other approaches (somewhat arbitrarily) choose only one representative. This is more consistent with the dense interconnected nature of regulatory circuits that control cell states. Second, Ontogenet allows us to choose a regulator in a context-specific manner, assuming that it may be relevant to the regulation of a gene module only in some cell types, but not others. Third, Ontogenet uses the lineage tree to guide its search for a regulatory program,

by preferring (but not mandating) models where ‘close’ cells in the lineage share regulatory mechanisms.

ImmGen regulatory model

The transcriptional response of the mouse hematopoietic system was separated into modules of co-expressed genes in two levels of resolution. At the lower resolution, 81 coarse-grained modules, including genes with broadly similar expression patterns were defined. Each coarse-grained module was further separated into fine-grained modules, representing smaller groups of genes with more coherent and tighter expression patterns, resulting in 334 fine-grained modules. For example, the coarse-grained module C33 is induced in B-cells, and separated into fine-grained modules that are induced in different subsets of B-cells, or at different progenitors of B-cells (Box2 figure panel b). Ontogenet reconstructed a regulatory model to each of these modules. The regulatory models included 554 of the 580 candidate regulators, at least 175 (32%) of which have not been previously been implicated in the hematopoietic context they are predicted to regulate. Some known regulators were not identified by Ontogenet, in most cases because they are expressed at low levels (biologically or due to suboptimal probe sets design) and were filtered out.

Gene network architecture properties responsible for the diversity of hematopoietic cell types and their function

One of the major goals of the ImmGen phase 1 has been to define and sort operationally discrete cell subsets within a defined functional lineage as a systemic baseline measurement of transcriptome complexity. Compound perturbations of the system followed by iterative transcriptome samplings (phase 2) will in principle yield the complete dynamic range of the system transcriptome and all dominant regulators. This task is to a large degree constrained by the availability of reagents (antibodies to cell surface markers or engineered cell type-specific reporters) to segregate live cell subsets from the whole population as well as by the size of the cell subset.

So far most of ImmGen data is generated from adult (4-6 wk old) male C57BL/6 mice. At the last release of the first phase of ImmGen (June 2012), the dataset comprises 816 arrays of 249 cell types, sorted from 27 tissues in 11 labs. All cell types were harvested under standard procedures, and the array hybridization was done in a central facility, in order to ensure minimal experimental biases. The lineage relationships between the cell types are described in Fig. 1b. The lymphoid branch of the lineage tree is much more structured and includes more known progenitors, whereas the myeloid branch has few known progenitors, and many cell types with no known relationships between them. Thus, currently, the lymphoid system is comparatively information-rich for studying transcriptional evolution along differentiation, whereas the myeloid lineage is ripe for identifying novel precursor-progeny and inter-subset lineage and functional relationships.

The phase 1 ImmGen studies can be broadly divided into three categories of survey: (1) gene expression profiles across hematopoietic subsets[10-14]; (2) developmental intermediates[12, 13, 15, 16]; and (3) subset-specific perturbations[10, 15, 17]. Here we

highlight the major themes that have emerged in each category and readers are directed to the published reports for more details.

Transcriptome-based lineage relationships in the hematopoietic system

Myeloid versus lymphoid transcriptional program—When the two major branches of the hematopoietic system, lymphoid and myeloid, are compared, gene modules and regulators most tightly linked to each can be discerned (Fig. 2). Many of the pan-myeloid regulators with strongest activity are known regulators, such as Batf3, Cebpa, Cebpb, Mafb, and Relb[18]. The shared modules reflect not only the properties of the pioneering progenitors of the lineage but also common functions. For example, module C24 that is induced in all myeloid cells is predicted to be activated in all myeloid cells by Ctbp2, Creg1, Tcfe3, Cebpa, Xbp1, and Bach1. C24 contains complement components, genes with anti-viral activity, and tissue remodeling genes.

The global analysis also reveals known regulators, previously implicated to have only a limited sphere of influence over the lineage, that may have a more broad activity. An example is Rbpj, a factor necessary for induction of Notch target genes, and previously shown to control CD11b+ dendritic cells (DCs) differentiation and function[19], but identified by Ontogenet as an activator of both CD11b+ and CD11b- DCs in the plasmacytoid DC-specific module F150. Panlymphoid modules are fewer (only 3 coarse-grained modules classified as lymphoid-specific, compared to 12 classified as myeloid specific), in agreement with the high degree of inter-lymphoid lineage divergence, compared to divergence in the myeloid lineage [9]. Ets1 is the only well-characterized regulator of lymphocytes identified in this manner.

Myeloid lineage subsets distinction and regulation—In the myeloid lineage, while subsets within the macrophage or DC sublineage show extensive transcriptome variability underpinning subset diversity at a resting state, the number of genes uniquely expressed in either sublineage was constrained, with the expression of 14 and 24 genes restricted to macrophages and classical DCs, respectively[11, 16]. This is very different from the inter-sublineage differences in the lymphoid lineage that can involve thousands of genes (e.g. B vs. T [20], γδ T vs. αβ T [12]). Core gene signatures unique to myeloid sublineages identify new markers (MerTK and CD64 and for macrophages; cKit and BTLA for DCs) that should aid in more pure separation of the cell subsets from other myeloid cells[11, 16]. Across tissues, any one macrophage subset is significantly more divergent in a tissue-restricted manner to other macrophages than any particular DC cell subsets sampled in the same way[11]. Moreover, all migratory DCs in tissue-draining lymph nodes share common gene signatures, accented by genes that may impose immune tolerance to self[16]. These results reinforce a pattern of strong influence of tissue environments and cell trafficking in specifying myeloid immune surveillance function, with most likely parallel environmental adaptations by lymphoid cell subsets, as illustrated by distinct gene expression profiles of T cell subsets in the gut in comparison to the pattern associated with corresponding populations in other anatomical locales[12, 13, 21].

Gata6 and Spi-C are examples of novel regulators of macrophages as identified by Ontogenet. While Gata6 is a regulator of macrophage-specific modules, it may be particularly important for peritoneal macrophages based on the expression pattern. Similarly, Spi-C may perform specialized function for splenic red pulp macrophages[11]. Fine modules characteristic of distinct DC subsets have common predicted regulators. Most of these regulators are known, such as Batf3, Irf8, and Ciita, but many more have not been characterized yet in the DC context, suggesting that the subset-specific regulators in concert with the pan-DC lineage factors induce transcriptional programs that define functionally distinct cell types[16].

Lymphoid lineage subsets distinction—Within the thymus, several subsets of T cells arise with diverse effector or regulatory functions [12, 22, 23]. Historically, T cells have been segregated by the type of TCR expressed (γδ or. αβ TCR) or by function (helper/regulatory or cytotoxic). Module analysis reveals a different property for lineage separations. Overall, thymocyte subsets can be clustered based on innate (functionally active within hours or a couple of days post infection or environmental alterations, with preprogrammed immune effector repertoires, and no significant difference between initial and subsequent responses) or adaptive (builds up to maximal responses within several days with acquired effector functions driven by pathogen elicited milieu, and subsequent response faster and stronger than initial response) classifications rather than the type of TCR expressed. For example, γδ T cells expressing Vδ6.3 TCR are intrathymically programmed to secrete both IL-4 and IFNγ by inducing Zbtb16 (PLZF), a TF exclusively expressed in αβ T cells with innate-like function[23]. The prototype of PLZF+ aβTCR+ cell types is the invariant αβ NKT cells expressing the canonical Vα14 TCR. Vδ6.3+ and Vα14+ T cell subsets have a significant overlap in transcriptomes[12, 13], likely accounting for their common functional capabilities and potential functional redundancies [24]. The transcriptional program induced in αβ iNKT cells is made up of elements of innate NK cells and adaptive αβ T cells[13]. Together, these results raise the possibility that the primary transcriptome-based subset stratification in lymphocytes is dictated by distinct molecular programs associated with innate versus adaptive immune function.

Gene network architectures during hematopoietic cell differentiation

Gene network modeling is capable of identifying novel candidate regulators for modules induced in specific cell types, as well as modules and regulators repeatedly used during development and in fully differentiated cell states. An intriguing example for cell type-specific module and its novel candidate regulators is module C60, induced only in fetal liver stem and progenitor cells, but not in adult bone marrow stem and progenitor cells. C60 contains members of the H19 network (Igf2, H19, and Dlk1) implicated in the control of fetal and postnatal growth in mice, and Plag1, a known regulator of the H19 network[25]. Ontogenet predicts Plag1 and Hmga2, a target of Lin28b[26], to be regulators of C60. Lin28b has been shown to be a determining factor of fetal lymphopoiesis[27] and induces the expression of genes, such as Zbtb16, necessary for the generation of innate lymphocytes, which appear to originate preferentially during fetal gestations or early postnatally[27-29].

As a prototype of repeatedly utilized regulators, TCF-1 (encoded by *Tcf7)*, a T cell lineage specification factor[30, 31], possesses far-ranging activities in diverse cell types (for examples, intrathymic γδ effector cell subset differentiation[32], control of NK cell receptor expression[33], maintenance of CD4+ Th17 cells[34], and generation of CD8+ memory αβ T cells[35]. Tcf7 is identified as a regulator in two coarse-grained modules and 18 fine-grained modules. TCF-1 controls diverse biological processes by extensive interactions with, and modulations by, co-factors that themselves are often cell type-restricted (e.g. Sox13 in γδ T cells). Predictably, many of the regulators that specify cell lineage fate, such as Irf4, Foxp3, and Rorγt, are characterized by complex interactomes as the driving force of their finely tuned functional activities[2, 5, 36].

The detailed cellular characterization of developmental intermediates in lymphoid cell sublineages offers an ideal system to probe the dynamic molecular basis of cell lineage specification. As a framework, gene transcription dynamics associated with hematopoietic cell differentiation are generally modeled on three sequential processes: First, a rapid and transient onset of activities of transcription factors responsible for turning on genes associated with differentiated states in immature lineage precursors; second, suppression of genes associated with the immediate precursor state or alternate cell fate choice; and third, long lasting induction of genes associated with differentiated states[12, 15, 37-40]. In the thymus, early thymic progenitors generate innate γδ T and adaptive αβ T cells. The general model applies well to the innate T cell development where the lineage specific TFs (e.g. Sox13, Etv5, Zbtb16, and Tox2) are turned on very early in differentiation and are shut off upon maturation into specialized effectors in the thymus[12]. A similar force acts on innate-like αβ iNKT differentiation where key early regulators (e.g. Zbtb16) are downmodulated upon maturation[13]. In contrast, in adaptive T cell differentiation the apex regulators do not achieve maximal induction until well after T cell lineage commitment and their expression is maintained in mature T cell subsets that exit the thymus. However, many of these regulators are turned off after T cell activation and differentiation into specialized effector subsets[15]. Hence, both innate and adaptive T cells follow the general dynamics of TF controlled lineage differentiation, the major difference being that for innate T cells, the differentiation window is compressed and mostly restricted to the thymus, whereas for adaptive T cells, their differentiation is marked by more than one wave of core regulator expression, and is not fully terminated until the memory phase. In this sense, the transcriptome dynamics in development fully accounts for the “fast” and “slow” nature of innate and adaptive lymphocytes, and provide the molecular blueprint for the widely held view that the thymus exports “memory-like” innate T cell subsets.

Temporal evolution of lymphoid regulatory modules post pathogen encounters

Initial infection studies involving innate (NK), innate-like (NKT) and adaptive (CD4+ and CD8+ αβ T cells) lymphocytes have provided preliminary leads to the molecular rewiring involved in the generation of short-lived effectors and long-lasting memory cells of the lymphoid lineage[10, 13, 17].

A hallmark of the adaptive T cell response is the massive expansion of pathogen-specific T cells within days of infection. Predictably, antigen-specific CD8+ T cells responding to

Listeria monocytogenes infection maximally induce transcription of a large cluster of genes within 48 hr, with the majority involved in proliferation, activation (TCR and cytokine-mediated), and increased metabolism. The memory phase (>45 days post-infection) involves reversing much of the initial super-induction and expression of a limited number (13) of memory-specific genes, including Bcl2 and Cdh1[17]. Ontogenet analysis identified known players in effector and memory CD8+ T cell subset differentiation (e.g., Tbx21, Erg2, Egr3, Prdm1, and Tcf7), as well as candidate regulators of short-term effector/memory cells, such as Rora, Tox, and Zeb2. Interestingly, CD8+ T cell gene clusters associated with the memory phase were operational in a subset of innate-like NKT cells and activated γδ T cells[17], suggesting that similar functions in different cell types are programmed by common regulators (e.g., fine-grained module F99). Conversely, innate NK cells (independent of virus encounters) and intestinal intraepithelial γδ T cells were most similar to effector CD8+ T cells. Thus, the preprogrammed innate lymphoid subsets as a group possess a full spectrum of adaptive effector and memory CD8+ regulatory programs, reinforcing the theme that cell function is a central determinant of the transcriptome, and by extension, cell lineage relatedness. Note however that how each cell type attains particular functional specialization can be unique. This array of unique responses likely minimizes the impact of variants that can escape canonical immune responses.

Insights from integrating ImmGen with other datasets

The power of ImmGen dataset is not only in the analysis of the data within it, as described above, but in the integration with external systemic datasets, which can amplify informational outputs and yield new paradigms. Here, we describe as an example comparing ImmGen with a similar, though much more limited, human dataset[40]. The recently published ENCODE data[41], which aims to map the functional elements in the human genome, is another prominent candidate for integration.

A recent study, termed differentiation map or ‘D-MAP’, collected an expression compendium of 39 cell types (211 samples) from human immune and hematopoietic lineages[40]. Comparison of ImmGen and D-MAP expression profiles show that most orthologous genes between human and mouse show similar expression patterns in the immune lineages, though there are some differences[42].

Concluding remarks

The scope, uniform data collection and quantitation procedures, and centralized regulatory model construction of the ImmGen compendium have established the baseline measurement of variations in the hematopoietic transcriptomes that allow for many novel analyses. There remain some gaps in the survey (e.g. fetal hematopoietic system, non-lymphoid tissue-resident hematopoietic cell types). In many ways, the completion of phase I is the starting point for unraveling the molecular circuits dictating the versatility and fitness of mammalian immunity. This effort will necessitate incorporating new methods from other fields (Box 2), and combining transcriptome dynamics with global chromatin analyses (epigenetic and core TF occupancy) across resting and perturbed hematopoietic cell subsets. The resultant context-dependent inducible gene modules will provide testable leads to lineage-specific

mechanistic studies of cell development and function as well as novel, network-integrated drug targets for hematopoietic diseases. Given some notable shortcomings of mouse as a model for human immunology and disease processes[43], a community effort to build a human ImmGen seems warranted to fully exploit the information explosion that is expected to follow during the next phases of ImmGen.

Acknowledgments

We thank the ImmGen labs, L. Lanier, S. Itzkovitz, A. Elbaz and S. Tal for discussion, L. Gaffney for help with the figures, and eBioscience, Affymetrix, and Expression Analysis for support of the ImmGen Project. Supported by NIH CA100382, AI101301 to J.K., and AI072073 to ImmGen. The **ImmGen Project Consortium consists of:** Paul Monach, Susan A Shinton, Richard R Hardy, Radu Jianu, David Koller, Jim Collins, Roi Gazit, Brian S Garrison, Derrick J Rossi, Kavitha Narayan, Katelyn Sylvia, Joonsoo Kang, Anne Fletcher, Kutlu Elpek, Angelique Bellemare-Pelletier, Deepali Malhotra, Shannon Turley, Adam J Best, Jamie Knell, Ananda Goldrath, Vladimir Jojic, Daphne Koller, Tal Shay, Aviv Regev, Nadia Cohen, Patrick Brennan, Michael Brenner, Taras Kreslavsky, Natalie A Bezman, Joseph C Sun, Charlie C Kim, Lewis L Lanier, Jennifer Miller, Brian Brown, Miriam Merad, Emmanuel L Gautier, Claudia Jakubzick, Gwendalyn J Randolph, Francis Kim, Tata Nageswara Rao, Amy Wagers, Tracy Heng, Michio Painter, Jeffrey Ericson, Scott Davis, Ayla Ergun, Michael Mingueneau, Diane Mathis, and Christophe Benoist.

References

1. Lee Y, et al. Induction and molecular signature of pathogenic TH17 cells. Nat Immunol. 2012; 13:991–999. [PubMed: 22961052]
2. Ciofani M, et al. A Validated Regulatory Network for Th17 Cell Specification. Cell. 2012; 151:289–303. [PubMed: 23021777]
3. Amit I, et al. Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. Science. 2009; 326:257–263. [PubMed: 19729616]
4. Mullen, Alan C., et al. Master Transcription Factors Determine Cell-Type-Specific Responses to TGF-β Signaling. Cell. 2011; 147:565–576. [PubMed: 22036565]
5. Glasmacher E, et al. A Genomic Regulatory Element That Directs Assembly and Function of Immune-Specific AP-1–IRF Complexes. Science. 2012; 338:975–980. [PubMed: 22983707]
6. Benoist C, et al. Consortium biology in immunology: the perspective from the Immunological Genome Project. Nat Rev Immunol. 2012; 12:734–740. [PubMed: 22955842]
7. Hijikata A, et al. Construction of an open-access database that integrates cross-reference information from the transcriptome and proteome of immune cells. Bioinformatics. 2007; 23:2934–2941. [PubMed: 17893089]
8. Heng TSP, et al. The Immunological Genome Project: networks of gene expression in immune cells. Nat Immunol. 2008; 9:1091–1094. [PubMed: 18800157]
9. Jojic V, et al. Identification of transcriptional regulators in the mouse immune system. Nat Immunol. 2013 accepted.
10. Bezman NA, et al. Molecular definition of the identity and activation of natural killer cells. Nat Immunol. 2012; 13:1000–1009. [PubMed: 22902830]
11. Gautier EL, et al. Gene-expression profiles and transcriptional regulatory pathways that underlie the identity and diversity of mouse tissue macrophages. Nat Immunol. 2012; 13:1118–1128. [PubMed: 23023392]
12. Narayan K, et al. Intrathymic programming of effector fates in three molecularly distinct [gamma] [delta] T cell subtypes. Nat Immunol. 2012; 13:511–518. [PubMed: 22473038]
13. Cohen NR, et al. Shared and distinct transcriptional programs underlie the hybrid nature of iNKT cells. Nat Immunol. 2012 advance online publication.
14. Malhotra D, et al. Transcriptional profiling of stroma from inflamed and resting lymph nodes defines immunological hallmarks. Nat Immunol. 2012; 13:499–510. [PubMed: 22466668]
15. Mingueneau M, et al. The transcriptional landscape of αβ-T cell differentiation. Nat Immunol. 2013 accepted.

16. Miller JC, et al. Deciphering the transcriptional network of the dendritic cell lineage. Nat Immunol. 2012; 13:888–899. [PubMed: 22797772]
17. Best JA, et al. Transcriptional insights into the CD8+ T cell response to infection and memory T cell formation. Nat Immunol. 2013 in press.
18. Satpathy AT, et al. Re(de)fining the dendritic cell lineage. Nat Immunol. 2012; 13:1145–1154. [PubMed: 23160217]
19. Lewis, Kanako L., et al. Notch2 Receptor Signaling Controls Functional Differentiation of Dendritic Cells in the Spleen and Intestine. Immunity. 2011; 35:780–791. [PubMed: 22018469]
20. Painter MW, et al. Transcriptomes of the B and T Lineages Compared by Multiplatform Microarray Profiling. The Journal of Immunology. 2011; 186:3047–3057. [PubMed: 21307297]
21. Shires J, et al. Biological Insights into TCRγδ+ and TCRαβ+ Intraepithelial Lymphocytes Provided by Serial Analysis of Gene Expression (SAGE). Immunity. 2001; 15:419–434. [PubMed: 11567632]
22. Rothenberg EV, et al. Launching the T-cell-lineage developmental programme. Nat Rev Immunol. 2008; 8:9–21. [PubMed: 18097446]
23. Godfrey DI, et al. Raising the NKT cell family. Nat Immunol. 2010; 11:197–206. [PubMed: 20139988]
24. Pereira P, Boucontet L. Innate NKTγδ and NKTαβ cells exert similar functions and compete for a thymic niche. European Journal of Immunology. 2012; 42:1272–1281. [PubMed: 22539299]
25. Varrault A, et al. Zac1 Regulates an Imprinted Gene Network Critically Involved in the Control of Embryonic Growth. Developmental Cell. 2006; 11:711–722. [PubMed: 17084362]
26. Kiel MJ, et al. Spatial differences in hematopoiesis but not in stem cells indicate a lack of regional patterning in definitive hematopoietic stem cells. Developmental Biology. 2005; 283:29–39. [PubMed: 15913595]
27. Yuan J, et al. Lin28b Reprograms Adult Bone Marrow Hematopoietic Progenitors to Mediate Fetal-Like Lymphopoiesis. Science. 2012; 335:1195–1200. [PubMed: 22345399]
28. Grigoriadou K, et al. Most IL-4-Producing γδ Thymocytes of Adult Mice Originate from Fetal Precursors. The Journal of Immunology. 2003; 171:2413–2420. [PubMed: 12928388]
29. Haas, Jan D., et al. Development of Interleukin-17-Producing γδ T Cells Is Restricted to a Functional Embryonic Wave. Immunity. 2012; 37:48–59. [PubMed: 22770884]
30. Weber BN, et al. A critical role for TCF-1 in T-lineage specification and differentiation. Nature. 2011; 476:63–68. [PubMed: 21814277]
31. Germar K, et al. T-cell factor 1 is a gatekeeper for T-cell specification in response to Notch signaling. Proceedings of the National Academy of Sciences. 2011; 108:20060–20065.
32. Malhotra N, et al. A network of High Mobility Group box transcription factors programs innate IL-17 production. Immunity. 2013 in press.
33. Held W, et al. Clonal Acquisition of the Ly49A NK Cell Receptor Is Dependent on the trans-Acting Factor TCF-1. Immunity. 1999; 11:433–442. [PubMed: 10549625]
34. Muranski P, et al. Th17 Cells Are Long Lived and Retain a Stem Cell-like Molecular Signature. Immunity. 2011; 35:972–985. [PubMed: 22177921]
35. Jeannet G, et al. Essential role of the Wnt pathway effector Tcf-1 for the establishment of functional CD8 T cell memory. Proceedings of the National Academy of Sciences. 2010; 107:9777–9782.
36. Fu W, et al. A multiply redundant genetic switch 'locks in' the transcriptional signature of regulatory T cells. Nat Immunol. 2012; 13:972–980. [PubMed: 22961053]
37. Miyamoto T, et al. Myeloid or Lymphoid Promiscuity as a Critical Step in Hematopoietic Lineage Commitment. Developmental Cell. 2002; 3:137–147. [PubMed: 12110174]
38. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. Nat Genet. 2009; 41:553–562. [PubMed: 19377474]
39. Amit I, et al. A module of negative feedback regulators defines growth factor signaling. Nat Genet. 2007; 39:503–512. [PubMed: 17322878]
40. Novershtern N, et al. Densely Interconnected Transcriptional Circuits Control Cell States in Human Hematopoiesis. Cell. 2011; 144:296–309. [PubMed: 21241896]

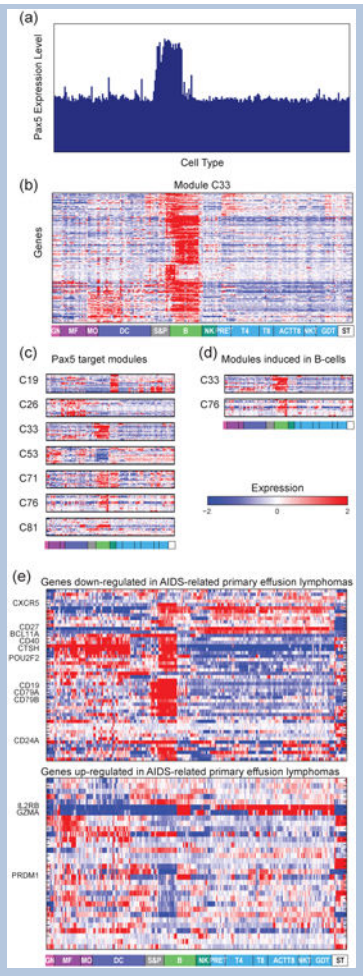
41. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]
42. Shay T, et al. Conservation and divergence in the transcriptional programs of the human and mouse immune systems. Proceedings of the National Academy of Sciences. 2013; 110:2946–2951.
43. Davis MM. A prescription for human immunology. Immunity. 2008; 29:835–838. [PubMed: 19100694]
44. Klein U, et al. Gene expression profile analysis of AIDS-related primary effusion lymphoma (PEL) suggests a plasmablastic derivation and identifies PEL-specific transcripts. Blood. 2003; 101:4115–4121. [PubMed: 12531789]
45. Shannon, CE.; Weaver, W. The mathematical theory of communication. University of Illinois Press; 1949.
46. Mossel E, Servedio RA. Learning juntas. Proc 35th Ann ACM Symp on the Theory of Computing. 2003:206–212.
47. Bendall SC, et al. Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. Science. 2011; 332:687–696. [PubMed: 21551058]
48. Mayr H, Ofial AR. The Reactivity–Selectivity Principle: An Imperishable Myth in Organic Chemistry. Angewandte Chemie International Edition. 2006; 45:1844–1854.

Text box 1

Usages of ImmGen

We envision several applications of ImmGen by the community, most of which are easily accomplished by one of the ImmGen browsers (http://www.immgen.org/):

1. Determining immune cell type(s) in which a gene of interest is expressed, using the Skyline viewer (e.g. Pax5, **Fig. a**).
2. Identifying the module to which a gene of interest is assigned to, that is which genes are co-expressed with a gene of interest, using the Modules viewer (e.g. C33, the Pax5 module, presented as a heatmap, **Fig. b**). This will also give information about the predicted regulators for this gene's module, by Ontogenet, enriched sequence motifs, and binding events.
3. Tracking the modules predicted to be controlled by a regulator of interest (e.g., Pax5, **Fig. c**).
4. Identifying differentially expressed genes between groups of immune cell types, using the Population Comparison browser.
5. Browsing modules induced in specific cell types (e.g., B-cells, **Fig. d**).
6. Extracting the expression patterns of a group of genes of interest across all ImmGen celltypes. This is possible for several predefined groups of genes using the Gene families Browser. Although this application is not currently available for user defined groups, it can be performedusing custom made scripts (e.g., genes down-regulated (top) or up-regulated (bottom) in AIDS-related primary effusion lymphoma samples compared to other tumor subtypes and normal B lymphocytes[44], **Fig. e**).
7. Determining conservation of expression pattern of a specific gene or module in the human immune system using the human mouse comparison browser.



Text Box 1 Figure. Examples of ImmGen application for gene(s) focused studies

(**a**) Bar chart of the immune cell type(s) in which a Pax5 is expressed, using the Skyline viewer. (**b**) A heatmap of the gene expression of the genes in coarse-grained module C33. Each row is a gene, each column is a cell type. Colorbar below represents cell lineage. (**c**) Heatmap of coarse-grained modules predicted to be controlled by Pax5. Each row is a gene, each column is a cell type. Colorbar below represents cell lineage. (**d**) B-cells induce modules. Each row is a gene, each column is a cell type. Colorbar below represents cell lineage. (**e**) Heatmap of genes down-regulated (top) or up-regulated (bottom) in AIDS-related primary effusion lymphoma samples compared to other tumor subtypes and normal B lymphocytes[44], **Fig. e**). Only genes assigned to ImmGen modules are shown, sorted by module number. Selected gene symbols are shown to the left.

Text box 2

Applying methods from other disciplines to ImmGen dataset

The scope and structure of the ImmGen dataset invite application of computational methods not generally familiar to most immunologists. One of the enriching aspects of the consortium is that dialogues between groups of different scientific backgrounds can produce new questions and applications of methods not strictly related to gene expression analysis to provide possible answers. For example, can the information content of the transcriptome of each cell type in the system approximate its level of differentiation, more specifically the pluripotency of HSC vs. the restricted lineage potency of its progenies? Hematopoiesis is a process in which an HSC has all the information required to generate blood and immune cell types, whereas terminally differentiated cells can only create the same cell type. HSC and terminally differentiated cells carry the same genetic information, but their transcriptome (and other parameters) is different, and likely responsible for the difference in the lineage potential. Information Theory is a mathematical field that studies the information content of sets of numbers[45], which can be applied to address the basis for the hierarchical nature of hematopoiesis.

As a second example, all modeling methods of transcriptional regulators assume that the transcriptome of a cell, which includes tens of thousands of genes, is a function of the activity of a limited set of regulators, typically dozens to hundreds of genes. The Junta problem is defined as finding a function whose output is dependent on only a constant number of input parameters, typically much smaller than the number of output parameters [46]. This is an open problem with a lot of research in Machine Learning. Defining a regulatory model is equivalent to finding a function whose output is the transcriptome of a cell, and input is the expression of the regulators, making it a Junta problem The ImmGen dataset suggests a case study to apply heuristics developed for the Junta problem on a real life problem.

In the last example, cell types of the immune system are mostly defined by a somewhat arbitrarily selected set of the cell surface proteins they express. It was shown that there are intermediate cell types between well characterized cell types and lineages[47]. This can be interpreted as a result of the arbitrary definition, or of incompletely characterized cells. However, the cells may follow the chemical reactivity-selectivity principle[48], stating that the more reactive the molecule, the less selective the response. A system optimized for speed must trade off on the selectivity, typically by producing intermediate outputs, or have an output production pathway that is not specific. The functional cell should be able to generate a myriad of outputs to potentially variable inputs (e.g. pathogens). Hence, applications of this principle to the ImmGen dataset can produce testable hypotheses.

Highlights

- ImmGen is a public resource to study genes and their networks in the immune system
- ImmGen datasets are used to gain insights into cell lineage relatedness and origins
- A regulatory computational model identifies known and novel regulatory interactions
- Web-based ImmGen data browsers are portals to multivariate genomic data analysis

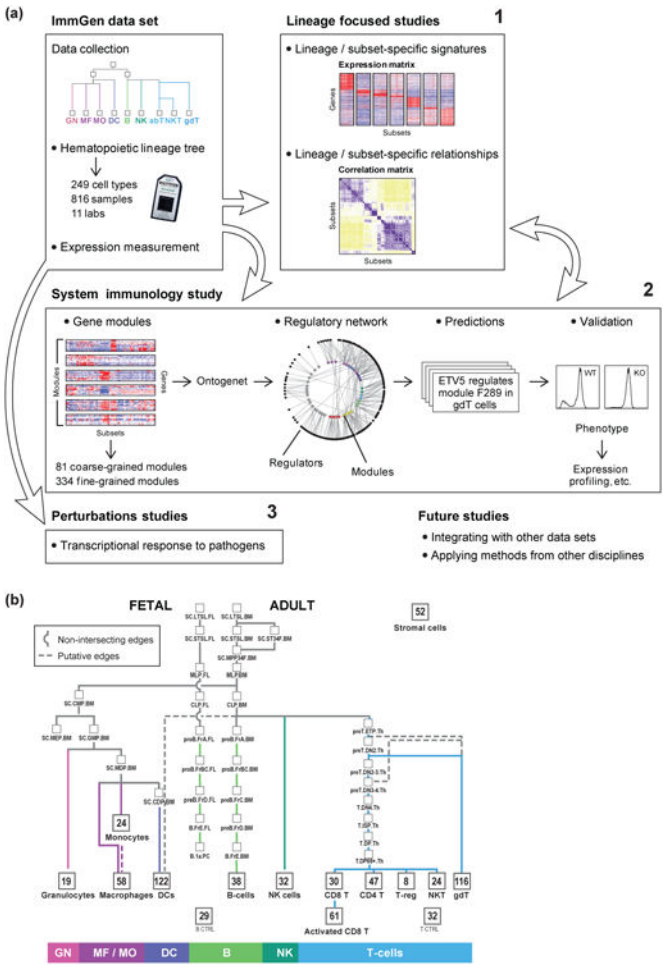


Figure 1. The ImmGen compendium and its analysis

(a) ImmGen dataset (top, left), and the three types of studies performed by the ImmGen consortium: (1) Lineage focused studies that identified lineage specific genes (e.g. heatmaps of gene expression across cell subsets), described the relationship between lineages (represented by Pearson correlation matrix) and assigned cell types to the lineages tree. (2) System studies that defined modules of coexpressed genes across the entire dataset and reconstructed the modules' regulatory program using Ontogenet. The resulting regulatory network described organizational principles governing the differentiation within the immune system, and identified candidate novel regulatory factors that can be experimentally verified. (3) Perturbations studies that investigated the transcriptional response to pathogens. ImmGen dataset can be integrated with other datasets, and methods from other disciplines can be applied to it, to gain new biological perspectives (Future studies). **(b)** The hematopoietic lineage tree constructed by ImmGen. Big rectangles with numbers represent a group of cell types, with the number of replicates listed. Small rectangles represent single cell type, typically with 3 replicates. Lines/edges between rectangles represent differentiation steps.

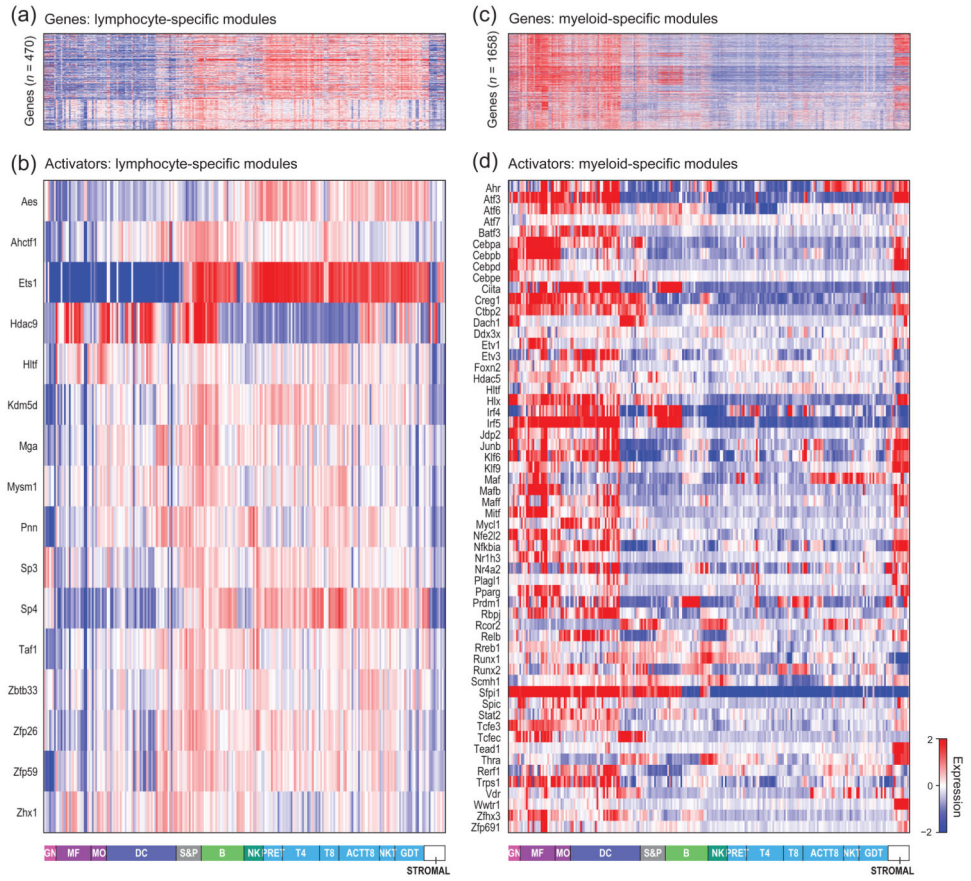


Figure 2. Lymphoid and myeloid specific modules and activators

Conserved pan-regulators of lymphoid and myeloid branches can be gleaned from the comparison of lymphoid and myeloid-specific modules. **(a)** Expression matrix of the genes in the modules that are induced in myeloid cells (Coarse modules 24, 25, 28, 29, 30, 31, 32, 45, 49, 74, 75 and 77). **(b)** Expression matrix of the activators assigned to the modules that are induced in myeloid cells with a maximal activity weight > 0.05 (arbitrary threshold). **(c)** Expression matrix of the genes in the modules that are induced in lymphoid cells (Coarse modules 16, 21 and 22). **(d)** Expression matrix of the activators assigned to the modules that are induced in lymphoid cells with a maximal activity weight > 0.5. Blue-red color bars show relative expression level

Table 1

data browsers of the ImmGen dataset, available at www.ImmGen.com

Browser name	Purpose
Gene skyline	Display expression profile of a single gene in a group of cell types
Gene families	Display interactive heatmaps of predefined gene families
Gene expression map	Color coded expression of genes along the genome
Population comparison	Find differentially expressed genes between user defined groups of cells
Modules and regulators	Browsing modules of co-expressed genes, their annotation and predicted regulators
RNA-seq	Visualize RNA-seq data of B and CD4+ T cells
Human mouse comparison	Display expression profile of a single gene in seven cell types measured in human and mouse, the expression matrix of the module this gene was assigned to in the specified species, and the expression of the corresponding orthologous genes in the other species.