HHS Public Access

Author manuscript

Nat Immunol. Author manuscript; available in PMC 2013 December 01.

Published in final edited form as:

Nat Immunol. 2013 June ; 14(6): 633–643. doi:10.1038/ni.2587.

Identification of transcriptional regulators in the mouse immune system

Vladimir Jojic[1,*], **Tal Shay**[2,*], **Katelyn Sylvia**[3], **Or Zuk**[2], **Xin Sun**[4], **Joonsoo Kang**[3], **Aviv Regev**[2,5,‡], **Daphne Koller**[1,‡], and **the Immunological Genome Project Consoritum**

[1]Computer Science Department, Stanford University, 353 Serra Mall, Stanford California 94305-9010, USA

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

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

[4]Laboratory of Genetics, University of Wisconsin-Madison, 425 Henry Mall, Madison, Wisconsin 53706, USA

[5]Howard Hughes Medical Institute, Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA.

Abstract

The differentiation of hematopoietic stem cells into immune cells has been extensively studied in mammals, but the transcriptional circuitry controlling it is still only partially understood. Here, the Immunological Genome Project gene expression profiles across mouse immune lineages allowed us to systematically analyze these circuits. Using a computational algorithm called Ontogenet, we uncovered differentiation-stage specific regulators of mouse hematopoiesis, identifying many known hematopoietic regulators, and 175 new candidate regulators, their target genes, and the cell

‡to whom correspondence should be addressed: aregev@broad.mit.edu (AR), koller@cs.stanford.edu (DK). .
*These authors contributed equally

ImmGen Project Consortium: Paul Monach[6], Susan A Shinton[7], Richard R Hardy[7], Radu Jianu[8], David Koller[8], Jim Collins[9], Roi Gazit[10], Brian S Garrison[10], Derrick J Rossi[10], Kavitha Narayan[3], Katelyn Sylvia[3], Joonsoo Kang[3], Anne Fletcher[11], Kutlu Elpek[11], Angelique Bellemare-Pelletier[11], Deepali Malhotra[11], Shannon Turley[11], Adam J Best[12], Jamie Knell[12], Ananda Goldrath[12], Vladimir Jojic[1], Daphne Koller[1], Tal Shay[2], Aviv Regev[2], Nadia Cohen[13], Patrick Brennan[13], Michael Brenner[13], Taras Kreslavsky[11], Natalie A Bezman[14], Joseph C Sun[14], Charlie C Kim[14], Lewis L Lanier[14], Jennifer Miller[15], Brian Brown[15], Miriam Merad[15], Emmanuel L Gautier[15,16], Claudia Jakubzick[15], Gwendalyn J Randolph[15,16], Francis Kim[17], Tata Nageswara Rao[17], Amy Wagers[17], Tracy Heng[18], Michio Painter[18], Jeffrey Ericson[18], Scott Davis[18], Ayla Ergun[18], Michael Mingueneau[18], Diane Mathis[18] & Christophe Benoist[18]

[6]Department of Medicine, Boston University, Boston, Massachusetts, USA. [7]Fox Chase Cancer Center, Philadelphia, Pennsylvania, USA. [8]Computer Science Department, Brown University, Providence, Rhode Island, USA. [9]Department of Biomedical Engineering, Howard Hughes Medical Institute, Boston University, Boston, Massachusetts, USA. [10]Immune Diseases Institute, Children's Hospital, Boston, Massachusetts, USA. [11]Dana-Farber Cancer Institute and Harvard Medical School, Boston, Massachusetts, USA. [12]Division of Biological Sciences, University of California San Diego, La Jolla, California, USA. [13]Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Boston, Massachusetts, USA. [14]Department of Microbiology & Immunology, University of California San Francisco, San Francisco, California, USA. [15]Icahn Medical Institute, Mount Sinai Hospital, New York, New York, USA. [16]Department of Pathology & Immunology, Washington University, St. Louis, Missouri, USA. [17]Joslin Diabetes Center, Boston, Massachusetts, USA. [18]Division of Immunology, Department of Microbiology & Immunobiology, Harvard Medical School, Boston, Massachusetts, USA.

types in which they act. Among the novel regulators, we highlight the role of ETV5 in γδT cells differntiation. Since the transcriptional program of human and mouse cells is highly conserved[1], it is likely that many lessons learned from the mouse model apply to humans.

Introduction

The Immunological Genome Project (ImmGen) is a consortium of immunologists and computational biologists who aim, using shared and rigorously controlled data generation pipelines, to exhaustively chart gene expression profiles and their underlying regulatory networks in the mouse immune system[2]. In this context, we provide the first comprehensive analysis of the ImmGen compendium, using a novel computational algorithm to reconstruct a modular model of the regulatory program of mouse hematopoiesis.

Understanding the regulatory mechanisms underlying the differentiation of immune cells has important implications for the study of development and for understanding the basis of human immune disorders and hematologic malignancies. Most studies of hematopoiesis view differentiation as a process controlled by relatively few 'master' transcription factors (TFs), expressed in specific lineages that act to set and reinforce distinct cell states[3]. However, a recent analysis of gene expression in 38 cell types in human hematopoiesis[4] suggested a more complex organization involving a larger number of transcription factors that control combinations of modules of co-expressed genes and are arranged in densely interconnected circuits. The human study was restricted, however, to human cells that could be obtained in sufficient quantities from peripheral or cord blood and thus could not access many immune cell populations. The 246 mouse immune cell types in the 816 arrays of the ImmGen compendium offer an unprecedented opportunity to study the regulatory organization of hematopoiesis within the context of a rich and diverse lineage tree. Since the transcriptional program of human and mouse cells is highly conserved[1], it is likely that many lessons learned from the mouse model will be applicable in humans.

Two key approaches to identify regulatory networks[5] are physical models based on the association of a TF or a *cis*-regulatory element with a target's promoter (*e.g.*, from ChIP-Seq), and observational models that infer regulation from a statistical dependence between the level or activity of a TF (at the protein or mRNA level) and that of its presumed target. In both cases, analyzing the relationship between a putative regulator and a module of co-regulated targets enhances robustness and biological interpretability[6-7]. Physical data provides direct evidence for biochemical interactions but is not necessarily functional[8] and is challenging[9] to collect, whereas mRNA profiles are highly accessible but provide only correlative evidence. Since physical and observational models are complementary, using both[4] can enhance our confidence[5-7] and expand our scope of discovery.

Analyzing cells organized in a known lineage, as in hematopoiesis, offers unique opportunities that have not been leveraged before. In particular, previous models[4] did not explicitly consider the fact that more closely related cells (according to the known lineage tree) likely share many of their regulatory mechanisms, and that regulatory relations that exist in one sub-lineage may not be active in another. Incorporating this information may help identify true regulators of hematopoiesis .

Here, we use these insights to develop a novel computational method, Ontogenet, and apply it to the ImmGen compendium to build an observational model associating 578 candidate regulators with modules of co-expressed genes. Modules were defined at two levels of resolution – with 81 larger coarse-grained modules, some of which were further refined into smaller modules with more coherent expression, resulting in 334 fine-grained modules The model identifies many of the known hematopoietic regulators, was supported using a complementary physical model, and proposed dozens of new candidate regulators. Our model provides a rich resource of testable hypotheses for experimental studies, and the Ontogenet algorithm can be used to decipher regulation within the context of any cell lineage.

Results

A transcriptional compendium of mouse immune cells

The ImmGen consortium dataset[2] (the April 2012 release) consists of 816 expression profiles from 246 mouse immune cell types (Fig. 1, Supplementary Table 1). The cell types span all major hematopoietic lineages, including stem and progenitor cells (S&P), granulocytes, monocytes, macrophages, dendritic cells (DC), natural killer (NK) cells, B cells, and T cells. T cells include many αβ T cells types, regulatory T cells (T_{regs}), natural killer T cells (NKT) and γδ T cells. The ‘same’ cell type was often sampled from several tissues, such as bone marrow, thymus, and spleen.

Similarities in global profiles trace the cell ontogeny

Correlations in global profiles between samples are largely consistent with the known lineage tree (Fig. 2). In general, the closer two cell populations are in the lineage tree, the more similar their expression profiles (Pearson r=−0.71, Supplementary Fig. 1). Within myeloid cells, profiles are overall similar, with granulocytes the least variable, dendritic cells the most variable (consistent with their sampling from diverse tissues and their known inherent diversity[10]), and all myeloid cells weakly similar to stromal cells. Conversely, lymphocytes show larger differences between lineages. NK cells, while tightly correlated, do show weaker similarity to T cells, especially $CD8^+$ T cells and natural killer T cells. T cells are very heterogeneous, partly reflecting the finer sampling in this lineage. Stem cells are most similar to early myeloid and lymphoid progenitors (S&P group, Fig. 2), followed by pre-B and pre-T cells, consistent with a gradual loss of differentiation potential. As a resource for studying each lineage, we used one way ANOVA to define characteristic signatures of over- and under-expressed genes for each of the main eleven lineages, compared to all other lineages (Supplementary Table 2).

Coarse and fine-grained expression modules in hematopoiesis

To characterize the key patterns of gene regulation, we next defined modules of co-expressed genes, at two levels of granularity (Supplementary Fig. 2a, b). We first constructed 81 coarse-grained modules (C1-C81, Supplementary Fig. 2c-h, Supplementary Table 3), and then further identified for each coarse-grained module a set of nested fine modules (Supplementary Fig. 2a), resulting in 334 fine modules spanning 7,965 genes (F1-F334, Supplementary Table 4). Coarse modules help us capture the mechanisms that co-

regulate a larger set of genes in one lineage, whereas fine modules may help identify distinct regulatory mechanism controlling only a smaller subset of these genes in the other lineage(s). Many of the modules are enriched for coherent functional annotations, *cis*-regulatory elements (Supplementary Table 5), and binding of transcription factors (Supplementary Table 6, Supplementary Note 1), including with binding sites for factors known to act as regulators at the lineage(s) in which the module's genes are expressed (Supplementary Note 2). All modules and their associated enrichments can be searched, browsed and downloaded in the ImmGen portal (http://www.immgen.org/ModsRegs/modules.html).

Most coarse-grained modules (48 of 81 modules, 4,478 of 7,965 genes) show either lineage-specific induction (Supplementary Fig. 2c and 3) or pan-differentiation regulation (Supplementary Fig. 2d, e, 4 and 5). In addition, 6 modules are 'mixed use' across lineages (Supplementary Fig. 2f and 6), 8 are stromal specific (Supplementary Fig. 2g), and 19 display expression patterns that do not fall into these categories (Supplementary Fig. 2h and 7). Lineage-specific repression is rare (only in C53 – B cells, C17 – stromal cells).

Ontogenet: Reconstructing lineage-sensitive regulation

We next devised a new algorithm, Ontogenet, to decipher the regulatory circuits that drive hematopoietic cell differentiation. Ontogenet aims to fulfill several biological considerations: criterion 1, the expression of each module of genes is determined by a combination of activating and repressing transcription factors; criterion 2, the activity level of these factors may change in different cell types; for example, a factor A may activate a module in one lineage but not in another, even if A is expressed in both lineages; criterion 3, the identity and activity of the factors regulating a module are more similar between cells that are close to each other in the lineage tree (*e.g.*, from the same sub-lineage) than between 'distant' cells (*e.g.*, from two different sub-lineages), in accordance with the increased similarity in expression profiles between closer cell types (Supplementary Fig. 1); and criterion 4, lineage master regulators (*e.g.*, GATA3 for T cells) are active across the sub-lineages, but the sub-types can also have additional more specific regulators (*e.g.*, FOXP3 for T_{regs}). The former should be captured as shared regulators of a coarse module and its nested fine modules, whereas the latter only regulate particular fine modules.

Ontogenet receives as input gene expression module, the lineage tree, and the expression profiles of a pre-designated set of 'candidate regulators' (transcription factors, chromatin regulators, *etc.*). It then associates each module with a combination of regulators (criterion 1 above), where each regulator is assigned an 'activity weight' in each cell type indicating its activity as a regulator for that module in that cell (criterion 2 above). The regulator activity is at the protein level, but is inferred solely from transcript levels. Following the previously published Lirnet[6], a method for regulatory network reconstruction, the activity-weighted expression of the regulators is combined in a linear model to generate a prediction of the modules' gene expression in each cell type (Fig. 3). In this model, the expression of the module's genes in a given cell type is approximated by the linear sum of the regulators' expression in that cell type multiplied by each regulator's activity weight in that cell type. As a result, the model makes predictions such as: "In pre B cells, Module 1 is activated by

transcription factors A and B and repressed by factor C, whereas in B cells, factors A and C are no longer active (even if the factors are expressed), and Module 1 is activated by B and D.". Our model assumes that all the genes in the same module are regulated in the same way. This is essential for statistical robustness, although it comes at a cost of missing some gene-specific expression patterns. The fine modules let us examine subtler expression patterns shared by fewer genes, but are more susceptible to noise.

While Ontogenet reconstructs a potentially different regulatory program for each cell type, as reflected by cell-specific activity weights for each regulator, it is geared toward maintaining the same activity level across consecutive stages in differentiation (criterion 3). This is achieved by penalizing changes in the activity weights of the regulatory program between a cell type and its progenitor. The fine-grained modules derived from a coarse-grained module 'inherit' the same regulators and activity weights that were inferred for their coarse-grained module (while possibly gaining additional regulators, criterion 4). Altogether, we use an optimization approach that constructs an ensemble of regulatory programs that try to achieve several goals: each regulatory program explains as much of the gene expression variance in the module as possible; the regulatory programs remain as simple as possible; regulatory programs are consistent across related cell types in the ontogeny, and fine modules have similar regulators to those of the coarse modules to which they belong.

Notably, the approach we and others previously used to identify combinations of regulators (*e.g.*, linear regression regularized using the Elastic Net penalty[6,11]) assumed that the regulatory activity (and hence activity weight) is the same across all cell types. Thus, if a regulator was expressed at the same level in two different cells, it was deemed active to the same extent. This violates the known context-specificity of regulation in complex lineages. Conversely, allowing the algorithm to construct a separate regulatory program for each cell type independently is impractical and also ignores the expected similarity in gene regulation between related cell types within the lineage. Ontogenet solves this problem by leveraging the lineage tree when inferring the regulatory connections and their activity, such that the module's genes' are more likely to be regulated in a similar way in related cell types.

Ontogenet regulatory model for mouse hematopoiesis

We applied Ontogenet to the 81 coarse-grained and 334 fine modules, a lineage tree consisting of 195 cell types, and 580 candidate regulators. The Ontogenet model identified 1,417 regulatory relations (1091 activating, 317 repressing, nine mixed) between 81 coarse-grained modules and 480 unique regulators (*e.g.*, Fig. 4, Supplementary Fig. 8, Supplementary Table 5 and http://www.immgen.org/ModsRegs/modules.html). On average, there were 17 regulators per coarse-grained module, and three coarse-grained modules per regulator. As determined by cross-validation, Ontogenet constructs regulatory programs that are strictly better in predicting new and previously unseen expression data than those obtained by Elastic Net[6], a method that does not use the tree and has fixed activity weights (Supplementary Fig. 9, Supplementary Note 3).

In most cases (59%), a regulator's activity weights vary between different cell types ('high changing'), reflecting context-specific regulation (Supplementary Fig. 10). When we prune

regulatory interactions whose maximal effect (defined as the product of activity weight and expression) is low, we obtain a sparser network, in which pan-differentiation and lineage specific modules are mostly controlled by distinct regulators (Fig. 5), whereas mixed-use modules share regulators with modules in the other classes.

The regulatory model associating 334 fine modules and 554 regulators in 6,151 interactions had qualitatively similar patterns, except for having more regulators with mixed behavior (high changing on some modules and low changing on others), probably reflecting both the increased number of interactions, and the finer regulatory program (Supplementary Fig. 10, Supplementary Table 7, http://www.immgen.org/ModsRegs/modules.html). This rich regulatory model for mouse immune system differentiation identified many known regulatory interactions, and suggests new regulatory interactions in specific immune contexts.

Ontogenet predicts known hematopoietic regulatory interactions

Many of the regulatory interactions identified by Ontogenet were previously known, supporting the accuracy of our model. For example, within individual regulators, PU.1 (encoded by *Sfpi1*) was selected as a regulator of the myeloid and B cells module C25 (and 13 of its 15 fine modules); C/EBPα (encoded by *Cebpa*) regulates the myeloid modules C24, C30 and C74, the macrophage module C29, and many myeloid fine modules; C/EBPβ (encoded by *Cebpb*) regulates myeloid specific modules C25 and C30, and many myeloid fine modules; MAFB regulates the macrophage specific modules C29, F128 and F131; STAT1 regulates the interferon response module C52; TBX21 (T-bet) regulates the NK module C19 and NKT module F288, and CIITA regulates the antigen presenting cells module F136.

Furthermore, the combination of regulators associated with a single module is also consistent with known regulatory relations. For example, the B cell module C33 is regulated by the known B cell regulators PAX5, EBF1, POU2AF1 and SPIB (Fig. 4); the T cell module C18 (Supplementary Fig. 8) is regulated by the known T cell regulators BCL11B, GATA3, LEF1, TOX and TCF7; the γδ T cells module C56 is regulated by the known γδ T cells regulators ZBTB16 (PLZF), SOX13 and ID3, all also involved in NKT development and function; the more γδ specific fine module F289 is regulated by all of these as well as ETV5, not previously associated with γδ T cells (discussed below); the NKT module F188 is regulated by GATA3, TBX21 and ZBTB16, and fine modules F150 and F152, in which $CD8^+$ DC cells expression is higher than $CD4^+$ DCs expression, are regulated by IRF8 (but not IRF4), consistent with the known role of subset-selective expression IRF4 and IRF8 in DC commitment[12].

Ontogenet's predictions are also supported by their significant overlap with those based on enrichment of *cis*-regulatory motifs and ChIP-based binding profiles in the modules (Supplementary Tables 5 and 6), supporting a direct physical interaction between a regulator and the genes in the module with which it was associated by Ontogenet (Supplementary Table 8). For example, 27 of the associations between a regulator and a coarse module are supported by *cis*-regulatory motif enrichment (p-value = 2.6×10^{-5}, hyper geometric test for two groups; p-value < 10^{-5} permutation test), for example, the GATA2 motif in HSC

module C40 and the SFPI1 (PU.1) motif in myeloid module C25. ChIP profiles support 21 regulator-coarse module associations (p-value = 2.2×10^{-5}, hyper geometric test for two groups; p-value < 10^{-5} permutation test), such as the binding of C/EBPα and C/EBPβ in the myeloid module C24 and the binding of EBF1 in the B cell module C33.

While these overlaps are statistically significant, they nevertheless also indicate that most regulatory interactions are not supported by enriched known *cis*-regulatory motifs or available TF binding data, and vice versa. There are three reasons for this. First, scoring for binding sites and their enrichment is a process that is highly prone to false negatives; this is particularly likely to occur in much smaller fine modules. Second, the majority of regulators chosen by Ontogenet do not have a characterized binding motif (60% of regulators, 334 of 554) nor ChIP binding data in any cell type (90% of regulators, 497 of 554). Such regulators can only be nominated by an expression-based method, such as Ontogenet, and should not be considered as false positives of our method. Finally, in many cases when we do find an enrichment in a *cis*-regulatory element or binding profile for a TF A in module B (300/551 *cis*-regulatory interactions (54%); 52 of 90 ChIP based interactions (57%)), the TF (A) and its target module (B) show little or no correlation in expression (absolute Pearson $r < 0.5$). In some cases, this will be due to a factor that is not itself transcriptionally regulated (a real 'false negative' of Ontogenet), but in many others the factor likely controls these targets in another cell type not measured in our study (and hence is not in fact a false negative of Ontogenet).

A few of the known regulators of immune system differentiation[13] were not identified by the model, due to various reasons. TAL1 and BMI1 did not pass the initial filtering criteria, being only expressed in HSCs, and hence were not provided as input. GFI1 was not assigned as a regulator in stem and progenitor cells or granulocytes, because its expression is highest in pre-T, and only sparse and intermediate in stem and progenitor cells and granulocytes. E2A (encoded by *Tcf3*) was not identified as a T cell regulator, perhaps because it is not specifically expressed in T cells, and is in general lowly expressed, possibly due to a bad probeset. XBP1 was not identified as a B cell regulator, because it has relatively low expression in B cells in our arrays, and is more highly expressed in myeloid cells.

The re-discovery of known regulators lends support to the many novel regulatory interactions in the model. Of the 475 regulators that Ontogenet associated with lineage specific modules or pan-differentiation modules, at least 175 (37%) are completely novel in this context. Among those, for example, KLF12 is predicted as a regulator of the NK module C19, but was not previously associated with NK cell regulation. GATA6 is predicted as a regulator of the macrophage specific modules C31, C50 and C58, but was not previously associated with macrophages. This is in agreement with the significantly reduced number of granulocyte-macrophage colonies from embryoid bodies of GATA6 knock out mice[14]. Finally, ETV5 is predicted by the model to be a regulator of the γδ T cell modules F287 and F289, a novel role discussed below.

Context specific regulation underlies mixed-use modules

Context specific regulation, where the same set of genes are regulated by one set of regulators in the context of one lineage, and by another set of regulators in the context of

another lineage, was previously reported in selected cases, for example *Rag2* regulation by GATA3 in T cells and PAX5 in B cells[15]. Ontogenet's ability to recover different regulatory programs for the same module in different parts of the lineage tree can help decipher the regulatory mechanisms underlying 'mixed use' modules, expressed in more than one lineage. For example, module C70 is induced both in T_{regs} and some myeloid populations. Each activation event is associated with different regulators in our model: FOXP3 in $CD4^+$ T cells (itself a member of the module, although not expressed in the DC subsets) and PIAS3, HSF2 and INSM1 in DCs. In another example, fine-grained module F300 is independently induced in both mature B and T cells. While some of its regulators are themselves 'mixed-use' in both lineages, others are B cell specific (ZFP318, RFX5 and CIITA) or T cell specific (*e.g.*, EGR2).

Regulatory recruitment and rewiring during differentiation

The majority of regulatory relations identified by Ontogenet are dynamic, as reflected by the change in their associated activity weights during differentiation. This change provides a bird's eye view of the 'recruitment' and 'disposal' of regulators (Fig. 6a). To characterize this, for each cell type, we identified all the regulatory interactions whose activity weight changes (either increases or decreases) between that cell type and its immediate progenitor (Supplementary Table 9), and the unique regulators and modules involved in those interactions. In this way, we identified modules and regulators that are recruited and strengthened (activity weight increases compared to progenitor) or disposed and weakened (activity weight decreases compared to progenitor) at each differentiation step. Notably, recruitment (or disposal) of regulators does not necessarily mean that the regulators' expression changes, but that the model suggests that their regulatory activity has changed for this set of targets. For example, during the differentiation of $CD8^+$ T cells from the CLP, 61 regulatory interactions are recruited, involving 34 modules and 49 regulators, only 15 of which were previously associated with T cell differentiation. In particular, in the differentiation step from DN4 to ISP T cells, Ontogenet independently identified the previously reported involvement of MXD4, BATF and NFIL3, as well as newly identified RCBTB1, PIAS3 and ITGB3BP (Fig. 6b, c). In another example, during the differentiation step leading to NK cells, the NK module C19 was assigned the known NK regulators EOMES, and TBX21 as activators. Both EOMES and TBX21 were also recruited as repressors over this differentiation step in other modules. The differentiation step leading to T_{reg} recruits the T_{reg} module C70, and its known regulators FOXP3, as well as CREM, which was previously proposed as a T_{reg} cell regulator[16]. Notably, because HSCs have no parent in our model, regulators active in HSCs will only be noted when no longer used at later points (*e.g.*, HOXA7 and HOXA9 were no longer used as activators at the MLP stage). The first recruited activator is MEIS1, recruited in module C42 on the differentiation step leading to the MLP, and later no longer used in T cells, in agreement with the previously reported methylation and silencing of MEIS1 during differentiation towards T cells[17].

Ranking of lineage activators and repressors

The activity weights assigned for each regulator at each differentiation point allowed us to identify and rank regulators as lineage activators and repressors based on the entire model (Fig. 6d, Supplementary Table 10). In this way we correctly captured many known

regulators of each lineage among the top ranked activators. For example, our model associates MYC, MYCN, GATA2, and MEIS1 with S&P cells, BCL11B, TCF7 and GATA3 with αβ-T cells; POU2AF1, PAX5, EBF1 and SPIB with B cells; EOMES, TBX21 and SMAD3 with NK cells and GATA3 and ZBTB16 with NKT cells. In addition, the model makes many predictions of lineage regulators that were not previously associated with those lineages. For example: in S&P cells: HLF; in granulocytes: DACH1, recently reported to regulate cell cycle progression in myeloid cells[18], BACH1 and NFE2; in macrophages: CREG1; in DCs: ATF6, ETV3, SKIL, NR4A2 and NR4A3, previously shown to be induced in viral infected DCs[19-20]; in monocytes: POU2F2 (Oct2), previously reported to be up-regulated with macrophage differentiation[21], and KLF13, a regulator of B and T cells[22] that has a higher expression in monocytes; in B cells: ZFP318; and in NKs: ELF4 (Gm9907), previously shown to control the proliferation and homing of $CD8^+$ T cells[23]. Notably, while this pan-model analysis is useful, it can de-emphasize the contribution of important regulators captured by the model in a more nuanced way, for example as acting only during a limited window of differentiation, but not present in the mature stage. Those are captured by the recruitment and disposal analysis shown above (Fig. 6).

Finally, by counting the number of changes in activity weights that occur (across all regulators) at each differentiation step (edge) we can identify those differentiation points where regulation is rewired most substantially (Supplementary Fig. 11). For example, 19 regulators are recruited to coarse modules (activity weight increases from zero) at ETP, and 28 at Tgd.th, including the known T cell regulator GATA3, and the known γδ T regulators ID3 and SOX13 (Supplementary Fig. 11a). At CLP, four regulators are disposed of (activity weight reduces to zero) by coarse modules, including the HSC regulators HOXA7, HOXA9 and HOXB3. Eighteen regulators are disposed at preT.DN2, including GATA1, MYC and MYCN (Supplementary Fig. 11b). Overall, rewiring is more prominent at higher levels in the lineage than at lower (more differentiated) ones, though this may be partly due to the reduced power to detect changes at cell types that have no other cells differentiating from them (terminally differentiated, also called leaves in the tree). The individual differentiation steps that carry the largest number of activity weight changes are found in the small intestine DCs, thymus γδ T cells, liver/lung DCs and preT.DN2.th, suggesting substantial regulatory rewiring in these cells, possibly due to tissue-specific effects. The regulatory model for fine modules identifies a larger number of regulatory changes (82% of differentiation steps change activity weight, compared to 65% for the coarse-grained module model), in particular in differentiation steps leading to ‘leaves’ (terminally differentiated cells; 67% vs. 48%). Thus, the fine-grained modules help uncover more cell type-specific regulation.

ETV5 is a novel regulator of γδ T cell differentiation

To test one of the model’s predictions *in vivo*, we centered on regulatory activators of lineage-specific modules with previously no known function in that lineage. A practical criterion was that the gene can be manipulated *in vivo* in a cell type-restricted manner. We focused on the Ets family member ETV5’s predicted role as a regulator of γδ T cell differentiation in modules F287 and F289 since its expression is highly restricted to the γδ T cell lineage. Although the model assigns several regulators to these modules, only two – SOX13 and ETV5 – are specific to the γδ T cell lineage. Both are expressed in distinct

thymic precursors, raising the possibility that they are among the earliest determinants of the lineage. SOX13 is a known regulator of γδ T cells, but ETV5 has not been implicated in γδ T cell development.

To test the regulatory role of ETV5 in γδ T cells, we assessed γδ T cell development and function in mice lacking ETV5 specifically in T cells (CD2p-*Cre*Tg+*Etv5*$^{fl/fl}$). As γδ TCR^{+} thymocytes transit from immature CD24 (HSA)^{hi} cells to mature CD24^{lo} cells, they acquire effector functions[24]. ETV5 is expressed at the highest level in γδ thymocytes expressing the Vγ2 TCR chain that constitute nearly half of all γδ T cells in postnatal mice. The majority of Vγ2^{+} cells differentiate into IL-17-producing γδ effector cells in the thymus[24]. Thus, one prediction of the model was that intrathymic IL-17-producing γδ effector cell development would be particularly impaired in the absence of ETV5. In the T cell specific ETV5 conditional knockout (CKO) mice the overall number of γδ T cells generated is comparable to control mice (Fig. 7a). However, there is a specific loss of mature Vγ2^{+} thymocytes (Fig. 7b, **top**). This may be due to inefficient activation, as indicated by the decreased expression of CD44, the nominal marker of lymphocyte activation, on Vγ2^{+} thymocytes, and a corresponding increase in CD62L expression, a marker of a naive state (Fig. 7b, **bottom**). Moreover, the residual mature thymocytes are impaired in the generation of IL-17-producing γδ effector cells (Fig. 7c). Mature Vγ2^{+} thymocytes from CKO mice have decreased expression of the transcription factor RORγt that induces *Il17* transcription, and both thymic and peripheral γδ T cells are impaired in the generation of CCR6^{+}CD27^{-}IL-17-producing γδ effector cells. These results support the prediction of our model and demonstrate that Etv5 is essential for proper intrathymic activation and maturation of the IL-17-producing γδ effector cell subset.

Studying the Ontogenet model on the ImmGen portal

To facilitate exploration and testing of other predictions of our model, we provide the full set of modules and regulatory model as part of the ImmGen portal, with relevant tools for searching, browsing and visually inspecting the results. Specifically, the ‘Modules and Regulators’ data browser on the ImmGen portal (http://www.immgen.org/ModsRegs/modules.html) is the gateway to the Ontogenet regulatory model for ImmGen. It allows the user to browse coarse-grained or fine-grained modules by their number, pattern of expression, a gene they contain, a regulator that is predicted to regulate them or the cell type in which they are induced. For each module we present the expression of its genes and predicted regulators (each as a heatmap), the activity weights of each regulator in each cell, and the module’s mean expression projected on the lineage tree (as in Fig. 4a-d). The module page also links to a list of the genes in the module, the regulators that are members of the module, the regulators predicted to regulate the module, the regulators suggested by enrichment of *cis* motifs and binding events of the module genes, and functional enrichments of the module. Finally, we provide links to download a table with the assignment of all genes to coarse and fine modules, the regulatory program of all modules, and the Ontogenet code.

Discussion

The ImmGen Consortium dataset provides the most detailed and comprehensive view of the transcriptional behavior of any mammalian immune system, and arguably of any developmental cell differentiation process. We used these data to analyze the regulatory circuits underlying these processes, from global profiles, to modules, to the transcription factors that control them. The unique features of our novel algorithm, Ontogenet, allow us to uncover regulatory programs active at specific differentiation stages, and to follow them as they unfold and rewire.

Our analysis automatically re-discovers many of the known regulators and their correct function, suggests novel roles for at least 175 additional regulators, not previously associated with hematopoiesis, and identifies points in the lineage where regulators are recruited to control a specific gene program or lose their regulatory role. Our ability to automatically rediscover many known regulators at the appropriate developmental stage, and the significant correspondence between the predicted regulators, known functions, and *cis*-regulatory and ChIP-Seq enrichments supports the likely quality of our novel predictions. Among those, we experimentally test and validate a novel role for Etv5 in the differentiation of γδ T effector subset. On-going studies indicate that Etv5 regulates IL-17-producing γδ effector cell differentiation by selectively controlling the expression of genes in the F289 γδ lineage-specific module.

Ontogenet's rich model allows us both to predict the specific biological context at which regulation occurs, to generalize broad roles for regulators, and to identify global principles of the regulatory program. On the one hand, the ability to identify regulators that act only during specific windows helps detect "early" programming TFs, whose expression is shut off when cells transit to the mature stage. On the other hand, integrating across the model's predictions in an entire lineage helps uncover TFs important for the maintenance of lineage identity or function, such as those that directly regulate the expression of effector molecules. Finally, generalizing across multiple regulators, we can identify those points at which regulatory control rewires most substantially and the regulators controlling this rewiring.

As in all expression-based methods to predict regulation, Ontogenet cannot directly distinguish causal directionality. To avoid arbitrary resolution of this ambiguity, Ontogenet allows several regulators with similar expression profiles to be assigned together as regulators of a module. The dense interconnected circuits and extensive auto-regulation in other mammalian circuits controlling cell states[4, 25] suggests that these are likely to have functional roles, although some may be 'false positives'. Conversely, the activation of other functional regulators may not be reflected at their expression levels, and some may have been filtered by our stringent criteria (*e.g.*, *Tal1*, a known HSC regulator). These may be captured by our complementary analysis of enrichment of modules in *cis*-regulatory motifs and binding of regulators. Another challenge is posed by genes with very unique expression profiles that are assigned to modules with similar but distinct expression profiles (*e. g.*, *Rag1* and *Rag2* in coarse module 5). The inferred regulatory program is unlikely to hold for those genes.

A similar study of human hematopoiesis[4] suggested substantial mixed-use of modules between lineages, whereas the mouse compendium suggests that most modules are lineage specific. As we show in a companion manuscript[1], the global profiles, lineage specific signatures, and gene co-expression patterns are otherwise broadly conserved between human and mouse. One possible reason for the lesser extent of 'mixed-use' in the mouse program is that while the mouse dataset contains many more cell types, it does not include erythrocytes, megakayocytes, basophils and eosinophils, where many of the 'mixed-use' patterns were observed in humans[4]. Notably, many regulators are shared across lineages. In particular, some regulators are only active in one lineage at some modules, but are shared between lineages at other modules. For example, ATF6 is an activator in all lineages in the myeloid modules C25, C45 and C49, but is a T cell specific repressor in the T cell precursor module C57, and a T cell specific activator in the B cell module C71.

Ontogenet is applicable to other differentiation datasets, including fetal data or cancer, when using other predictors as candidate regulators (*e.g.*, genetic variants as in Lirnet[6]), when cells are measured at both resting and stimulated state, or for protein expression data (*e.g.*, single-cell, high-dimensional phosphoproteomic mass cytometry data[26]). In each case, the ability to share regulatory programs for related cell types or conditions can both enhance our power and help with biological interpretation. Notably, Ontogenet currently depends on a pre-constructed ontogeny. While much is known about the hematopoietic lineage, some parts remain unstructured (*e.g.*, all DCs in the myeloid lineage) and some progenitors are not known (*e.g.*, for γδ T cells or other innate-like lymphocytes). This reflects in part inherent lineage flexibility, whereby several cell types can differentiate into the same cell type, but in part just our current lack of knowledge of the particular progenitor of a given cell type. Novel methods would be required to construct an ontogeny automatically or to revise an existing one. In other cases, Ontogenet's output can be used to refine a topology, by identifying edges that do not correspond to any changes in regulatory programs and can be removed without disconnecting the lineage.

The ImmGen Compendium, coarse and fine grained modules, and the identified regulators and regulatory relations are all available for interactive searching and browsing and for download in the ImmGen Portal, and will provide an invaluable resource for future studies of the role of gene regulation in cell differentiation and immune disease.

Methods

Dataset

Mouse expression was measured on Affymetrix Mogen1 arrays (Affymetrix annotation version 31). Sorting strategies for the ImmGen populations can be found on the ImmGen website (http://immgen.org). Gene expression data are deposited in the Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) under accession no. GSE15907.

As the ImmGen dataset gradually grew from 2010 to 2012, clustering, regulatory program reconstruction, and final presentation were performed on three different ImmGen releases (September 2010, March 2011, and April 2012), while attempting to maximize backward compatibility as much as possible. The clusters and the regulatory program are from the

September 2010 and March 2011 releases, chosen to ensure consistency with the other ImmGen Report papers that refer to them. Clustering was performed on ImmGen release of September 2010, across 744 samples, 647 of which remained in the April 2012 release. Ontogenet was applied to the ImmGen release of March 2011, only to the data of the 676 samples (195 hematopoietic cell types) that were connected to the hematopoietic tree. Thus, we maintained the membership in clusters from the earlier analysis, but only used some of the samples to learn the regulatory program. The heatmaps presented in the paper display 755 samples (244 cell types), excluding control samples. For simplicity, there are only 720 samples presented on the full tree (210 cell types). Supplementary Table 1 lists all the samples in the last ImmGen release (April 2012), and states for each sample if it was used in generating the modules, regulatory program reconstruction, the presented heatmaps and tree. The Web resource is continuously updated

Data preprocessing

Expression data was normalized as part of the ImmGen pipeline by RMA. Data was log2 transformed. For gene symbols with more than one probeset on the array, only the probeset with the highest mean expression was retained. Of those, only probesets with a standard deviation higher than 0.5 across the entire dataset were used for the clustering, resulting with 7,965 unique differentially expressed genes in the September 2011 release and 8,431 in the April 2012 release.

Lineage specific signatures

We calculated signatures for 11 lineages: GN, MF, MO, DC, B, NK, T4, T8, NKT, GDT and S&P. Assignment of samples into lineages is listed in Supplementary Table 2. One way ANOVA was performed for each of the 6,997 genes that have an expression value above $\log_2(120)$ in at least one lineage, followed by a *post hoc* analysis (Matlab functions anova1 and multcompare). For each of the 11 lineages, a gene was considered induced if it has significantly higher expression in that lineage compared to all other lineages. A gene was considered repressed if it has significantly lower expression in that lineage compared to all other lineages. FDR of 10% was applied to the ANOVA *p*-values of all genes.

Definition of modules

Modules were defined by clustering. For coarse-grained modules, clustering was performed by Super Paramagnetic Clustering[27] (SPC), a principled approach to choose stable clusters from a hierarchical setting. SPC was used because it does not require a pre-defined number of clusters, but identifies the number inherently supported by the data. The clusters defined by SPC are stable across a range of parameters, though they can display variable levels of compactness. SPC was run with default parameters, resulting in 80 stable clusters. Those are named coarse-grained modules C1-C80. The remaining unclustered genes were grouped into a separate cluster C81.

Each coarse-grained module was further partitioned to fine-grained modules by affinity propagation[28] clustering, with correlation as the affinity measure. The “self-responsibility” parameter that indicates the propensity of the algorithm to form a new cluster was set at 0.01. Affinity propagation was used because SPC and hierarchical clustering did not further

break the coarse modules. Affinity propagation could not be used for for clustering all genes, because it has to work with a sparsified affinity matrix.

Clustering resulted in 334 fine-grained modules, referred to in the text as fine modules F1-F334. On average, 3.9 fine-grained modules were nested in a single coarse-grained module. The minimal number of fine modules nested in a coarse-grained module was 1 (23 coarse-grained modules), and the maximal was 11 (7 coarse-grained modules).

Choice of candidate regulators

Candidate regulators were curated from the following sources: (**1**) The mouse orthologs of all the genes that were used as candidate regulators in a previous study of human hematopoiesis[4]; (**2**) genes annotated with the Gene Ontology term 'transcription factor activity' in mouse, human or rat; (**3**) genes for which there is a known DNA binding motif in TRANSFAC matrix database[29] v8.3, JASPAR[30] Version 2008 and experimentally determined position weight matrices (PWMs)[31-32]; and (**4**) genes with published ChIP-Seq or ChIP-chip data (Supplementary Table 11). Regulators that were not measured on the array or whose expression did not change sufficiently (standard deviation < 0.5 across the entire dataset) to be included in the clustering were removed, unless they were highly correlated (>0.85) with another regulator that passed the cutoff. This resulted in 578 candidate regulators (Supplementary Table 12).

Hematopoietic tree building

The hematopoietic tree (Fig. 1) was built by the ImmGen consortium members. Each group created its own sub-lineage tree, and the sub-lineage trees were connected based on the best current knowledge, though many edges are hypothetical (dashed lines, Fig. 1). There are two roots to the tree – long term stem cells from adult bone marrow (SC.LTSL.BM) and long term stem cells from fetal liver (SC.LTSL.FL). Each population is a node in the tree (square, Fig. 1). Edges indicate a differentiation step, an activation step, time (as in the activated T cells) or a general assumption of similarity in regulatory program (Supplementary Table 13). Some intermediate inferred nodes were added to group cell populations that are assumed to have a common progenitor or common regulatory program, but where this hypothetical population was not measured (*e.g.*, granulocyte and macrophages). For the populations that connected to more than one parent population, one of the edges was manually pruned, either the less likely one or arbitrarily, as listed in Supplementary Table 13.

Module regulatory program

Ontogenet takes as input (**1**) gene expression profiles across many different cell types, (**2**) a partitioning of the genes into modules (coarse-grained and fine-grained clusters, above); (**3**) a predefined set of candidate regulators; and (**4**) an ontogeny tree relating the cell types. It then constructs a regulatory program for each module consisting of a linear combination of regulators with possibly distinct 'activity weights' for each regulator in each cell type. A module regulatory program is the linear sum of the regulators expression multiplied by each regulator's activity weight, which approximates the expression pattern of the module. Each regulatory program aims to explain as much of the gene expression variance in the module as possible, while remaining as simple as possible and being consistent across related cell

types in the ontogeny. In a regular linear model, the activity weights are constant across all conditions. Here we allow a change of activity weights between cell types (Fig. 3).

Notably, all regulators are considered as potential regulators for each module. That includes regulators that are members of the module. Thus, a module can be assigned regulators that are its members, and regulators that are not its members, but regulators that are members of the module will not necessarily be assigned to it.

More formally, we model the expression of a gene in a module as a (noisy) linear combination of the expression of the regulators. We denote the activity of a regulator *r* in a cell type *t* as $a_{r,t}$. We model the expression of a gene *t*, a member of module *m*, in cell type *t* as $x_{i,t} = \sum_r w_{m,r,t} a_{r,t} + \in_{m,t}$, where each $\in_{m,t}$ is a Gaussian random variable with zero mean and variance $\sigma^2_{m,t}$ specific to a combination of a module *m* and a cell type *t*. Hence the regulatory program learned by Ontogenet is represented in terms of $w_{m,r,t}$ activity weights specific to a (module, regulator, cell type) combination. Due to parameter tying enforced by the model, the effective number of parameters is significantly smaller than the nominal size of the regulatory program representation (# modules) × (# regulators) × (# cell types).

Module cell-type specific variance estimation

The module variance in a given cell type $\sigma^2_{m,t}$ is estimated from the expression of the module's member genes across all replicates of the cell type. While we use an unbiased estimator, we make special considerations for the modules with less than 10 members. For these modules the variance estimate $\sigma^2_{m,t}$ is computed by a pooled variance estimator across modules with more than 10 members but still specific to the cell type. The estimated variances in a fine-grained module are typically smaller than the variances in its parent coarse-grained module.

Regulatory program fitting as a penalized regression problem

Estimation of the activity weights $w_{m,r,t}$ takes the form of a regression problem, but due to over-parameterization of the problem, we need to regularize it, using an extension of the fused Lasso framework[33], giving rise to a penalized regression problem of the form

$$\frac{1}{n_m}\sum_{i,t}\frac{1}{2\sigma^2_{m,t}}\left(x_{i,t}-\sum_r w_{m,r,t}\,a_{r,t}\right)^2+J\left(w\right),$$

where $J(w)$ is a chosen penalty. In our case, this penalty is composed of two parts, one promoting sparsity and selection of correlated predictors and another promoting consistency of regulatory programs between related cell types.

We assume that only a small number of regulators are actively regulating any one module. A standard approach to promoting such sparsity in regression problems is to introduce an L_1 penalty, the sum of absolute values $\sum_m\sum_r\sum_t|w_{m,r,t}|$. However, this penalty tends to be overly aggressive in inducing sparsity, thus pruning multiple highly correlated predictors and selecting only a single representative. This aggressive pruning may be inappropriate, since

the correlated regulators may all be biologically relevant due to 'redundancy' in densely interconnected regulatory circuits. Such behavior can be counteracted by the addition of squared terms $\frac{1}{2}\sum_m\sum_r\sum_t(w_{m,r,t})^2$ yielding a composite penalty known as Elastic Net[11] as previously proposed[6] $\lambda\sum_r\sum_t|w_{m,r,t}|+\frac{\kappa}{2}\sum_r\sum_t(w_{m,r,t})^2$, which we write compactly as $\lambda\|w_m\|_1+\frac{\kappa}{2}\|w_m\|_2^2$.

An important input to our regulatory program fitting procedure is the ontogeny (differentiation) tree (Supplementary Table 13). This tree is encoded as an edge list (f) and with $(t_1, t_2) \in f$ we denote that cell type t_1 is a parent of cell type t_2. The similarity of the regulatory programs for a particular module in two related cell types $(t_1, t_2) \in f$ can be assessed as a sum of the absolute value of the difference of activity weights in the two programs, $\sum_r|w_{m,r,t2} - w_{m,r,t2}|$. The key observation being that is $|w_{m,r,t2} - w_{m,r,t2}|$ is 0 if the regulatory relationship between regulator r and module m is the same in cell type t_2 and its parent type t_1. More generally, the total difference of the regulatory programs can be written as $\sum_{(t1,t2)\in f}\sum_r|w_{m,r,t1} - w_{m,r,t2}|$. We will write this term in a compact form as $\|Dw_m\|$ where w_m is a vector of activity weights for all regulators across all cell types concatenated together and D is a matrix of size (RE) × (RT), where R is the number of regulators, T is the number of cell types and E is the number of edges in the tree. We note that multiplication by the matrix D computes the differences between relevant entries of the vector w_m. The less the regulatory programs change throughout differentiation, the smaller the term $\|Dw_m\|$. Thus, using this term as a penalty will promote the preservation of a consistent regulatory program throughout differentiation.

Combining all the considerations above, the complete objective for fitting a regulatory program of a module m is given by

$$\frac{1}{n_m}\sum_{i,t}\frac{1}{2\sigma_{m,t}^2}\left(x_{i,t}-\sum_r w_{m,r,t}\,a_{r,t}\right)^2+\lambda\|w_m\|_1+\frac{\kappa}{2}\|w_m\|_2^2+\gamma\|Dw_m\|_1.$$

Optimization of this objective is somewhat complicated by the fact that absolute value is a non smooth function and hence direct optimization by methods such as gradient descent is not feasible, as these work only on smooth problems. Alternative methods, such as projected gradients, can be used, but their convergence is relatively slow. We therefore opted to use a primal dual interior point method[34]. Different choices of the parameters λ,κ,δ yield different regulatory models as solutions, with different data-fitting and model-complexity properties. We scanned sets of parameters in the range (The schedule for each of the parameters lambda, gamma and kappa was geometric: e^{-7}, e^{-6},…, e^3 spanning values between 0.001 and 20) and chose the optimal set of parameters using the Bayesian Information Criteria (BIC) (see 'Model selection using Bayesian Information Criterion' section).

In order to simplify the discussion of the optimization we introduce a sparse predictor matrix A of size (RT) × (T) where $A_{t,(r-1)T+t}=\frac{a_{r,t}}{\sigma_{mt}}$ and 0 otherwise. Further, we note that the

optimal w_m depends only on the mean expression profile of the module's genes and we can introduce variable $y_t = \frac{1}{\sigma_{mt}} \sum_{i \in m} \frac{1}{n_m} x_{i,t}$. Hence we can rewrite the objective as

$$\frac{1}{2}\|y - Aw_m\|_2^2 + \|w_m\|_2^2 + \lambda\|w_m\|_1 + \frac{\kappa}{2}\|w_m\|_2^2 + \gamma\|Dw_m\|_1.$$

Finally we can absorb the term $\frac{\mu}{2}\|w_m\|_2^2$ into the first term as follows

$$\frac{1}{2}\left\| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} A \\ \sqrt{\kappa} I \end{bmatrix} w_m \right\|_2^2 + \lambda\|w_m\|_1 + \gamma\|Dw_m\|_1.$$

Regulatory program transfer between coarse-grained and fine-grained modules

The fine-grained modules are encouraged to have a similar program to the coarse-grained module in which they are nested. This is accomplished by introduction of an additional penalty term. We will denote the already learned regulatory program of a coarse-grained module as w_0 and the regulatory program of a fine-grained module that we wish to learn as w_m. The coarse-to-fine version of our objective is then

$$\frac{1}{2}\|y - Aw_m\|_2^2 + \|w_m\|_2^2 + \lambda\|w_m\|_1 + \frac{\kappa}{2}\|w_m\|_2^2 + \gamma\|Dw\|_1 + \frac{\tau}{2}\|w_0 - w_m\|_2^2,$$

where the last term ties the coarse-grained and fine-grained modules' programs. This objective can be transformed into

$$\frac{1}{2}\left\| \begin{bmatrix} y \\ 0 \\ \sqrt{\tau} w_0 \end{bmatrix} - \begin{bmatrix} A \\ \sqrt{\kappa} I \\ \sqrt{\tau} I \end{bmatrix} w_m \right\|_2^2 + \lambda\|w_m\|_1 + \gamma\|Dw_m\|_1.$$

Solving the prototypical optimization problem

We note that both coarse-grained and fine-grained module regulatory program fitting problems have been expressed in the following general form

$$\underset{w}{\text{minimize}} \frac{1}{2}\|y - Xw\|_2^2 + \lambda\|w\|_1 + \gamma\|Dw\|_1.$$

We reformulate this optimization problem by addition of variables that decouple the penalties.

$$\begin{array}{rl} \text{minimize}_{\text{w,w}^1,\text{w}^2} & \frac{1}{2} r' r + \lambda\|z\|_1 + \gamma\|d\|_1 \\ \text{subject to} & r = y - Xw, z = w, d = Dw. \end{array}$$

This reformulation enables straightforward derivation of a primal dual interior point method[34].

Model selection using Bayesian Information Criterion

The formulation of our optimization problem above is dependent on a set of parameters λ,k,γ: we obtain a model by solving the convex problem above for a particular combination of λ,κ,γ. Different combinations of these parameters will yield regulatory programs of different quality. One way to identify the optimal λ,κ,γ, is by using held-out data or through cross validation. However, search for these parameters using cross-validation is prohibitively expensive. As an alternative, we use a model selection approach based on the Bayesian Information Criterion (BIC) to compare models resulting from different choices of these three parameters and select the best one. The BIC criterion compares models, here encoded by regulatory programs, based on their tradeoff between data log likelihood and degrees of freedom. The log likelihood for our model is

$$LL\left(w\right)=-\sum_{m}\sum_{i,t}\frac{1}{2\sigma_{m,t}^{2}}\left(x_{i,t}-\sum_{r}w_{m,r,t}\,a_{r,t}\right)^{2}+\text{const.}$$

The computation of the degrees of freedom is somewhat technically involved but intuitively simple: an activity weight that remains the same through a particular connected portion of the differentiation tree is counted as a single degree of freedom. In order to make this more formal we will consider matrix A and construct its counterpart B. We will use $A_{r,t}$ to denote a **column** of matrix A. We will now construct a graph where nodes correspond to columns of matrix A. Given two nodes corresponding to $A_{r,t1}$ and $A_{r,t1}$ the graph will have an edge between these two nodes if cell type t_1 is a parent of cell type t_2, and $w_{m,r,t1} = w_{m,r,t2}$. The matrix B will have columns that are sums of columns corresponding to connected components in the graph. We eliminate all columns of B that are zeros and the final degrees of freedom are given by $\mathrm{df}(w) = \mathrm{Trace}(B(B'B + \kappa\,\mathrm{diag}(c))^{-1}B')$ where $\mathrm{diag}(c)$ is a diagonal matrix with entries being a number of columns of A in the connected component associated with a column of B .

Hence we can compute the BIC(w) as

$$\begin{aligned}BIC\left(w\right)&=-2LL\left(w\right)+\mathrm{df}\left(w\right)log\,T\\&=\sum_{m}\sum_{i,t}\frac{1}{\sigma_{m,t}^{2}}\left(x_{i,t}-\sum_{r}w_{m,r,t}\,a_{r,t}\right)^{2}+\mathrm{df}\left(w\right)log\,T.\end{aligned}$$

Post-processing of regulatory programs

Once we obtain an optimal regulatory program with respect to BIC, we perform postprocessing to remove regulatory relationships for underexpressed regulators. We placed a low expression cutoff of 5.5 on the log2 scale. At this level the correlation between the predictor and the target module may very well be due to noise and hence the relationship could be spurious.

Systematic query for known functions of regualtors

For each lineage specific module, we automatically queried its regulators in Pubmed with the name of the lineage/s. For each module that is up or down-regulated with differentiation, we queried its regulators with 'hematopoietic differentiation'. All Pubmed queries were performed on July 30, 2012.

Choice of lineage regulators

For each lineage, we collected the regulators that were deemed active by having a non-zero activity weight and significant expression in excess of 9.0 on the log2 scale. For a given lineage, we deemed a regulator a lineage-activator if its average activity weight across all cell types in the lineage and all modules was positive. Analogously, a regulator was deemed a lineage-repressor if its average activity weight was negative. We subsequently ranked the regulators based on their average expression across cell types in which the regulator had a role. Hence, the regulators that were frequently active in a lineage and when active had higher levels of expression were ranked higher than the infrequently active or lowly expressed regulators. The regulators with the highest expression typically get the highest total activity weight across lineages.

Notably, this procedure – while straightforward – will not reflect all the lineage regulators identified by the model. First, those lineage regulators that act only during a limited window (e.g. early in differentiation) would be under-represented by this analysis, yet captured in the overall model in the window in which they act. Second, due to the postprocessing step (above) regulators with high baseline expression can have a constant activity weight even if their expression is very lineage specific (*e.g.* Gata3) and thus be under-represented in the recruitment analysis (though they too are chosen as regulators in the model).

Motif scanning

We scanned promoters of mouse genes for enriched motifs. We downloaded promoter sequences for mouse (mm9) from the UCSC Genome Browser website http://hgdownload.cse.ucsc.edu/downloads.html. For each gene, we scanned the region starting from −1,000 base-pairs (bp) upstream of the transcription start site (TSS), and ending at the +200 base-pairs downstream of the TSS. We represented the nucleotide at position j (relative to −1,000 bp from the TSS) for gene i as $S_{i,j}$. We represented each *cis*-regulatory element by a position weight matrix (PWM). We compiled a set of 1,651 PWMs from the TRANSFAC matrix database[29] v8.3, JASPAR[30] Version 2008, and experimentally determined PWMs[31-32]. For the k^{th} motif, we denote its PWM by P_k, a matrix of size $4\times L_k$ where L_k is the length of the motif, and $P_k(i,j)$ represents the probability of encountering the nucleotide j (j = 'A', 'C', 'G' or 'T') at the i^{th} position. For each gene i, a position along the promoter j, and a PWM k, we computed the local motif-matching score LOD(i,j,k), defined as the log-likelihood ratio (LOD score) for observing the sequence given the PWM versus a given random genomic background:

$$LOD(i,j,k) = \sum_{r=1}^{L_k} log_2 P_k(r, S_{i,j+r-1}) - log_2 P_b(S_{i,j+r-1})$$

Genomic background was determined as P_b('A') = P_b('T') = 0.3, P_b('C')=P_b('G') = 0.2, representing the nucleotide composition in the mouse genome. We then found the best motif instance over the entire promoter region, defined as *MAX-LOD(i,k) = max$_j$ LOD(i,j,k).*

Motif scoring threshold

We automatically computed a PWM-specific threshold by using the information content of each motif. The information content for the k^{th} motif is defined as,

$$IC_k = -\sum_{i=1}^{L}\sum_{j=1}^{4} P_k(i,j)\, log_2 P_k(i,j)$$

We defined the PWM-specific threshold for the k^{th} motif k as τ_k, the 1 - 2^{-IC_k} quantile of the PWM LODs distribution across all genes' promoters. We considered a 'hit' for the k^{th} motif at the i^{th} gene if the best score, *MAX-LOD(i,k)*, exceeded the threshold τ_k.

Motif enrichment in modules

For each module of genes *M*, and each motif *k*, we computed the *p*-value for enrichment, $p_e(M,k)$ of the motif in the module, compared to the entire set of genes assigned to modules serving as background. An enrichment of a motif in a module results in higher than expected *MAX-LOD* scores for the genes in this module – to capture this effect, we computed the *p*-value by comparing the scores *MAX-LOD(i,k)* for all genes *i* in the module *M* and the scores for the entire set of genes assigned to modules by performing a one-sided rank-sum test. We then employed an FDR of 5% on the entire matrix of *p*-values $p_e(M,k)$, and declared as significant hits all *p*-values passing this procedure. The FDR was calculated separately for coarse-grained and fine-grained modules.

Binding events enrichment

The public ChIP-Seq and ChIP-chip datasets listed in Supplementary Table 11 were downloaded (360 experiments of 109 unique regulators). The target list defined in each original publication was used whenever available. Otherwise, genes which had a binding event reported from the 1000 bp upstream to the TSS to the 200 bp downstream to the TSS were listed as targets. In datasets measured in human samples, gene symbols were replaced by the mouse gene symbol, whenever a one-to-one ortholog exists according to EnsemblCompara[35]. Only genes that were included in the clustering were considered as targets for the purpose of the calculation of enrichment.

Hypergeometric *p*-value was calculated for the size of intersection of each module with each target list. FDR of 10% was used for the entire table of *p*-values of all modules and all targets lists. The FDR was calculated separately for coarse-grained and fine-grained modules.

Estimating significance of regulatory program overlap

We report two *p*-values for each overlap of the three regulation models (from ChIP, *cis*-elements and Ontogenet). First, we calculated the hypergeometric test for two or three

groups, where the universe sizes are the number of possible regulatory interactions including the overlapping regulators. For example, to estimate the significance of the overlap of ChIP and Ontogenet regulatory interactions, the universe size is the number of regulators which were candidates to Ontogenet and have a ChIP information, times the number of modules. The ChIP interactions are the enriched modules per ChIP dataset, and the Ontogenet interactions are the regulators chosen for each module. Second, we calculated an empirical *p*-value from 10,000 permutations of the regulators in the regulatory interactions including the overlapping regulators. The latter *p*-values were calculated to account for the fact that some modules have more regulators than others. The hypergeometric *p*-values and the empirical *p*-values are similar for the overlap of each two methods, but different in significance for the three methods overlap, because the hypergeometric score for three groups explicitly takes into account the overlap between each two groups, whereas the empirical *p*-value does not.

Functional enrichment

The MSigDB version V.3 curated gene sets (C2), motif gene sets (C3) and Gene Ontology (GO) gene sets (C5) were downloaded from http://www.broadinstitute.org/gsea[36]. Positional gene sets (C1) for mouse were kindly provided by Arthur Liberzon, the MSigDB curator. For each group, gene symbols were replaced by the mouse gene symbol, whenever a one-to-one ortholog exists according to EnsemblCompara. Only genes that were included in the clustering were considered as functional group members for the purpose of the calculation of enrichment.

Hypergeometric *p*-value was calculated for the size of intersection of each module with each functional group. FDR of 10% was used for the entire table of *p*-values of all modules and all functional groups. The FDR was calculated separately for coarse-grained and fine-grained modules, and for the different classes of functional annotation (C1, C2, C3, C5).

Identification of differentiation steps with a change in activity weight of regulators

For each module and each edge (differentiation step) in the hematopoietic tree, the activity weight of the parent was compared to the activity weight of the child, resulting in one of several classifications: No change – activity weights are the same. Activator recruitment – parent activity weight is zero, child positive; Activator strengthening – parent activity weight positive and smaller than child; Activator disposal – parent activity weight positive and child zero. Repressor recruitment – parent activity weight is zero, child negative; Repressor strengthening – parent activity weight negative and larger than child; Repressor disposal – parent activity weight negative and child zero. For simplicity, we omitted the regulator weakening option. Note that those lineage specific regulators that are assigned constant activity weight across all cell types (*e.g.* Gata3) will not be captured by this analysis, but are part of the model.

Mice

Etv5$^{fl/fl}$ mice[37] were crossed to CD2 promoter driven Cre transgenic mice (C57BL/6) to generate conditional Etv5 knockout mice (CKO, CD2p-CreTg^{+}*Etv*$^{fl/fl}$, 3 times backcrossed to C57BL/6). The floxed locus is specifically deleted from the genome starting in

$CD25^{+}CD44^{-}CD3^{-}CD4^{-}CD8^{-}$ thymic precursors (termed DN3) with ~80% deletion efficiency in $\gamma\delta$ thymocyte subsets as inferred from the analysis of Cre activity reporter mice (CD2p-$CreTg^{+}$Rosa-$STOP^{fl/fl}$-EYFP). Five litters of neonates (5-7 days old) and adults (4 weeks old) containing $CreTg^{+}$ controls (CD2p-$CreTg^{+}$*Etv5*$^{+/+}$) and CKO mice were analyzed, each with 2-5 mice/genotype.

Flow cytometry

Intracellular (Cytofix/Cytoperm Kit, BD Biosciences) and intranuclear (FoxP3 Staining Kit, eBioscience) staining was performed as previously described[24]. The following molecules were detected using antibodies (Abs) purchased from eBioscience: Tcrδ (GL3), CD24 (HSA, M1/69), CD44 (IM7), CD62l (MEL-15), IL17A (ebio17B7) and RORγt (AFKJS-9). Abs specific for Vγ2 (UC3-10A6), Vδ6.3(8F4H7B7), CCR6 (140706), and CD27 (LG. 3A10) were purchased from BD Biosciences. Vγ1.1 Ab (2.11) was purified from culture supernatant and biotinylated using the FluoReporter Mini-Biotin-XX Labeling Kit (Invitrogen). Data were acquired on a BD LSRII cytometer and analyzed using FlowJo (Treestar).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Christophe Benoist, Roi Gazit, Paul A. Monach and all the members of the ImmGen Consortium for many helpful discussions, the ImmGen core team (M. Painter, J. Ericson, S. Davis) for help with data generation and processing, and eBioscience, Affymetrix, and Expression Analysis for support of the ImmGen Project. We thank Leslie Gaffney for help with figure preparation and layout of the lineage tree, Sigrid Hart for the initial layout of the lineage tree, and Arthur Liberzon for the positional gene sets for mouse. Work was supported by R24 AI072073 from NIH/NIAID to the ImmGen consortium, by an NIH Pioneer Award, a Burroughs Wellcome Fund CASI award, the Klarman Cell Observatory and HHMI (AR), and by NSF grant DBI-0345474, NIH U54-CA149145 and NIH 149644.0103 grants (VJ and DK). AR is a fellow of the Merkin Foundation for Stem Cell Research at the Broad Institute.

References

1. Shay T, Jojic V, Zuk O, Rothamel K, Puyraimond-Zemmour D, Feng T, Wakamatsu E, Benoist C, Koller D, Regev A, ImmGen Consortium. Conservation and divergence in the transcriptional programs of the human and mouse immune systems. Proceedings of the National Academy of Sciences. 2013; 110(8):2946–2951.
2. Heng TSP, Painter MW, Elpek K, Lukacs-Kornek V, Mauermann N, Turley SJ, Koller D, Kim FS, Wagers AJ, Asinovski N, Davis S, Fassett M, Feuerer M, Gray DHD, Haxhinasto S, Hill JA, Hyatt G, Laplace C, Leatherbee K, Mathis D, Benoist C, Jianu R, Laidlaw DH, Best JA, Knell J, Goldrath AW, Jarjoura J, Sun JC, Zhu Y, Lanier LL, Ergun A, Li Z, Collins JJ, Shinton SA, Hardy RR, Friedline R, Sylvia K, Kang J. The Immunological Genome Project: networks of gene expression in immune cells. Nat Immunol. 2008; 9(10):1091–1094. [PubMed: 18800157]
3. Iwasaki H, Akashi K. Myeloid Lineage Commitment from the Hematopoietic Stem Cell. Immunity. 2007; 26(6):726–740. [PubMed: 17582345]
4. Novershtern N, Subramanian A, Lawton LN, Mak RH, Haining WN, McConkey ME, Habib N, Yosef N, Chang CY, Shay T, Frampton GM, Drake ACB, Leskov I, Nilsson B, Preffer F, Dombkowski D, Evans JW, Liefeld T, Smutko JS, Chen J, Friedman N, Young RA, Golub TR, Regev A, Ebert BL. Densely Interconnected Transcriptional Circuits Control Cell States in Human Hematopoiesis. Cell. 2011; 144(2):296–309. [PubMed: 21241896]

5. Kim HD, Shay T, O'Shea EK, Regev A. Transcriptional Regulatory Circuits: Predicting Numbers from Alphabets. Science. 2009; 325(5939):429–432. [PubMed: 19628860]
6. Lee S-I, Dudley AM, Drubin D, Silver PA, Krogan NJ, Pe'er D, Koller D. Learning a Prior on Regulatory Potential from eQTL Data. PLoS Genet. 2009; 5(1):e1000358. [PubMed: 19180192]
7. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet. 2003; 34(2):166–176. [PubMed: 12740579]
8. Capaldi AP, Kaplan T, Liu Y, Habib N, Regev A, Friedman N, O'Shea EK. Structure and function of a transcriptional network activated by the MAPK Hog1. Nat Genet. 2008; 40(11):1300–1306. [PubMed: 18931682]
9. Ram O, Goren A, Amit I, Shoresh N, Yosef N, Ernst J, Kellis M, Gymrek M, Issner R, Coyne M, Durham T, Zhang X, Donaghey J, Epstein, Charles B, Regev A, Bernstein, Bradley E. Combinatorial Patterning of Chromatin Regulators Uncovered by Genome-wide Location Analysis in Human Cells. Cell. 2011; 147(7):1628–1639. [PubMed: 22196736]
10. Miller JC, Brown BD, Shay T, Gautier EL, Jojic V, Cohain A, Pandey G, Leboeuf M, Elpek KG, Helft J, Hashimoto D, Chow A, Price J, Greter M, Bogunovic M, Bellemare-Pelletier A, Frenette PS, Randolph GJ, Turley SJ, Merad M. Deciphering the transcriptional network of the dendritic cell lineage. Nat Immunol. 2012; 13(9):888–899. [PubMed: 22797772]
11. Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society. Series B (Statistical Methodology). 2005; 67(2):301–320.
12. Tamura T, Tailor P, Yamaoka K, Kong HJ, Tsujimura H, O'Shea JJ, Singh H, Ozato K. IFN Regulatory Factor-4 and -8 Govern Dendritic Cell Subset Development and Their Functional Diversity. The Journal of Immunology. 2005; 174(5):2573–2581. [PubMed: 15728463]
13. Orkin SH, Zon LI. SnapShot: Hematopoiesis. Cell. 2008; 132(4):712.e711–712.e712. [PubMed: 18295585]
14. Pierre M, Yoshimoto M, Huang L, Richardson M, Yoder MC. VEGF and IHH rescue definitive hematopoiesis in Gata-4 and Gata-6–deficient murine embryoid bodies. Experimental Hematology. 2009; 37(9):1038–1053. [PubMed: 19501129]
15. Kishi H, Wei X-C, Jin Z-X, Fujishiro Y, Nagata T, Matsuda T, Muraguchi A. Lineage-specific regulation of the murine RAG-2 promoter: GATA-3 in T cells and Pax-5 in B cells. Blood. 2000; 95(12):3845–3852. [PubMed: 10845919]
16. Bodor J, Fehervari Z, Diamond B, Sakaguchi S. ICER/CREM-mediated transcriptional attenuation of IL-2 and its role in suppression by regulatory T cells. European Journal of Immunology. 2007; 37(4):884–895. [PubMed: 17372992]
17. Ji H, Ehrlich LIR, Seita J, Murakami P, Doi A, Lindau P, Lee H, Aryee MJ, Irizarry RA, Kim K, Rossi DJ, Inlay MA, Serwold T, Karsunky H, Ho L, Daley GQ, Weissman IL, Feinberg AP. Comprehensive methylome map of lineage commitment from haematopoietic progenitors. Nature. 2010; 467(7313):338–342. [PubMed: 20720541]
18. Lee J-W, Kim H-S, Kim S, Hwang J, Kim YH, Lim GY, Sohn W-J, Yoon S-R, Kim J-Y, Park TS, Park KM, Ryoo ZY, Lee S. DACH1 regulates cell cycle progression of myeloid cells through the control of cyclin D, Cdk 4/6 and p21Cip1. Biochemical and Biophysical Research Communications. 2012; 420(1):91–95. [PubMed: 22405764]
19. Ng SSM, Chang T-H, Tailor P, Ozato K, Kino T. Virus-induced differential expression of nuclear receptors and coregulators in dendritic cells: Implication to interferon production. FEBS Letters. 2011; 585(9):1331–1337. [PubMed: 21492741]
20. Wang T, Jiang Q, Chan C, Gorski KS, McCadden E, Kardian D, Pardoll D, Whartenby KA. Inhibition of activation-induced death of dendritic cells and enhancement of vaccine efficacy via blockade of MINOR. Blood. 2009; 113(13):2906–2913. [PubMed: 19164597]
21. Neumann M, Fries H-W, Scheicher C, Keikavoussi P, Kolb-Mäurer A, Bröcker E-B, Serfling E, Kämpgen E. Differential expression of Rel/NF-κB and octamer factors is a hallmark of the generation and maturation of dendritic cells. Blood. 2000; 95(1):277–285. [PubMed: 10607713]
22. Outram SV, Gordon AR, Hager-Theodorides AL, Metcalfe J, Crompton T, Kemp P. KLF13 influences multiple stages of both B and T cell development. Cell Cycle. 2008; 7(13):2047–2055. [PubMed: 18604172]

23. Yamada T, Park CS, Mamonkin M, Lacorazza HD. Transcription factor ELF4 controls the proliferation and homing of CD8+ T cells via the Kruppel-like factors KLF4 and KLF2. Nat Immunol. 2009; 10(6):618–626. [PubMed: 19412182]
24. Narayan K, Sylvia KE, Malhotra N, Yin CC, Martens G, Vallerskog T, Kornfeld H, Xiong N, Cohen NR, Brenner MB, Berg LJ, Kang J. Intrathymic programming of effector fates in three molecularly distinct [gamma][delta] T cell subtypes. Nat Immunol. 2012; 13(5):511–518. [PubMed: 22473038]
25. Yosef N, Regev A. Impulse Control: Temporal Dynamics in Gene Transcription. Cell. 2011; 144(6):886–896. [PubMed: 21414481]
26. Bendall SC, Simonds EF, Qiu P, Amir E.-a.D. Krutzik PO, Finck R, Bruggner RV, Melamed R, Trejo A, Ornatsky OI, Balderas RS, Plevritis SK, Sachs K, Pe'er D, Tanner SD, Nolan GP. Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. Science. 2011; 332(6030):687–696. [PubMed: 21551058]
27. Blatt M, Wiseman S, Domany E. Superparamagnetic Clustering of Data. Physical Review Letters. 1996; 76(18):3251–3254. [PubMed: 10060920]
28. Frey BJ, Dueck D. Clustering by Passing Messages Between Data Points. Science. 2007; 315(5814):972–976. [PubMed: 17218491]
29. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E. TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. Nucleic Acids Research. 34(suppl 1):D108–D110. [PubMed: 16381825]
30. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Research. 2004; 32(suppl 1):D91–D94. [PubMed: 14681366]
31. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen X, Kuznetsov H, Wang C-F, Coburn D, Newburger DE, Morris Q, Hughes TR, Bulyk ML. Diversity and Complexity in DNA Recognition by Transcription Factors. Science. 2009; 324(5935):1720–1723. [PubMed: 19443739]
32. Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Peña-Castillo L, Alleyne TM, Mnaimneh S, Botvinnik OB, Chan ET, Khalid F, Zhang W, Newburger D, Jaeger SA, Morris QD, Bulyk ML, Hughes TR. Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences. Cell. 2008; 133(7):1266–1276. [PubMed: 18585359]
33. Tibshirani R, Saunders M, Saharon R, Zhu J, Knight K. Sparsity and Smoothness via the Fused Lasso. Journal of the Royal Statistical Society. Series B (Statistical Methodology). 2005; 67(1): 91–108.
34. Boyd, SP.; Vandenberghe, L. Convex optimization. Cambridge, Cambridge, UK ; New York: 2004.
35. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Research. 2009; 19(2): 327–335. [PubMed: 19029536]
36. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005; 102(43):15545–15550.
37. Zhang Z, Verheyden JM, Hassell JA, Sun X. FGF-Regulated Etv Genes Are Essential for Repressing Shh Expression in Mouse Limb Buds. Developmental Cell. 2009; 16(4):607–613. [PubMed: 19386269]

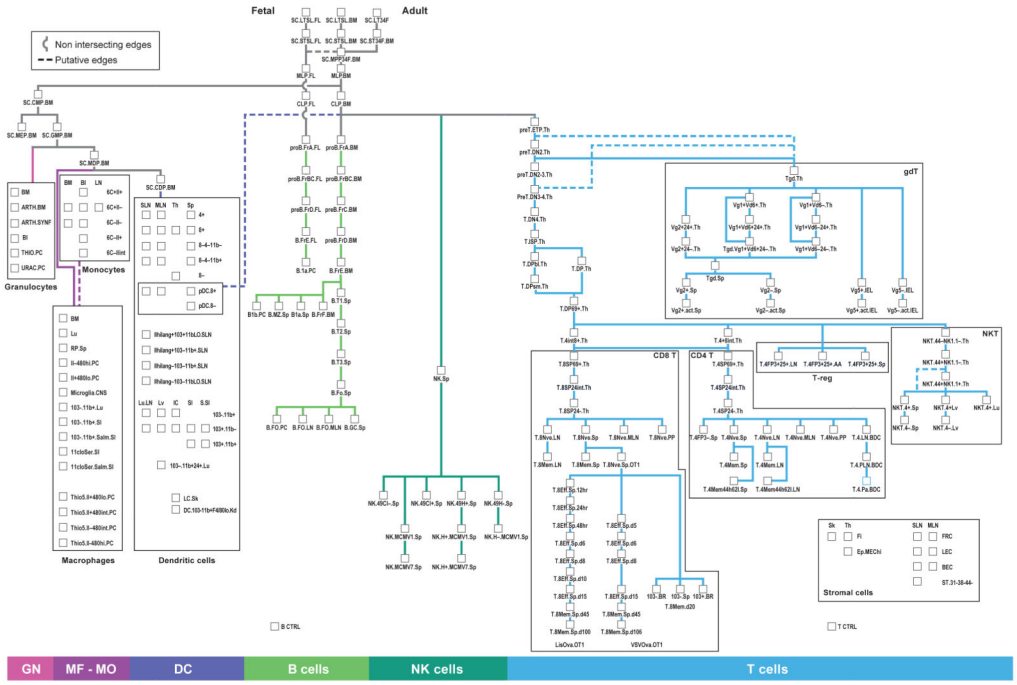

Figure 1.

Mouse cell populations in the ImmGen Compendium. Shown is the lineage tree of the hematopoietic mouse cell types profiled by the Immunological Genome consortium (www.immgen.org). Some stem, progenitor and B cells were sampled from adult and fetal liver. Stromal cells (box, bottom right) were also measured as part of ImmGen, but are not part of the lineage tree. The markers used to sort each cell population are listed (Supplementary Table 1). Adapted from [1].

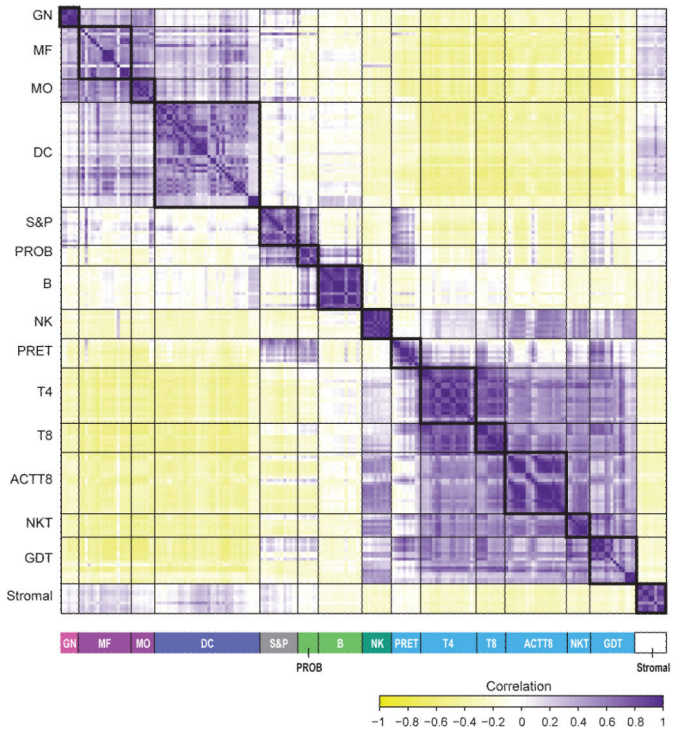

Figure 2.

Related cells have highly similar expression profiles. Shown are the Pearson correlation coefficients (purple – positive correlation; yellow – negative correlation; white – no correlation) between each pair of profiled cell types, calculated across the 1,000 genes (of the 8431 unique expressed genes) with the highest standard deviation across all samples. Black lines delineate major lineages. GN – granulocytes, MF – macrophages, MO – monocytes, DC – dendritic cells, S&P – stem and progenitor cells, PROB – preB and proB cells, NK – natural killer cells, T4 – $CD4^+$ T cells, T8 – $CD8^+$ T cells, ACTT8 – activated $CD8^+$ T cells, NKT – natural killer T cells, GDT – gamma delta T cells. Samples are sorted by breadth-first search on the tree in Fig. 1, with stromal cells at the lower or right end.

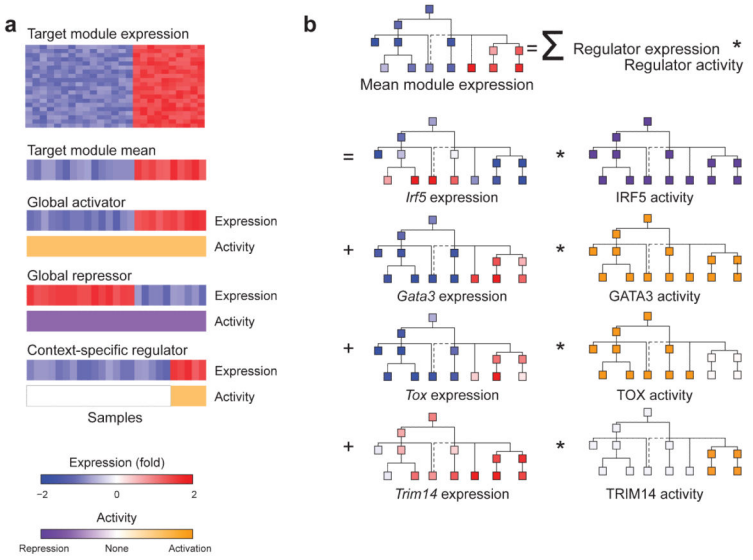

Figure 3.

Overview of Ontogenet. (**a**) Different types of regulators (bottom) can explain the expression of a module (top). For each regulator, we display its expression profile (blue/red vectors) and activity profile (orange/purple vectors). A regulator may have a uniform positive activity weight across the lineage (constitutive activator, top), a uniform negative activity weight (constitutive repressor, middle), or variable activity weights (context-specific regulator, bottom). (**b**) The mean expression of a module (top) is a linear combination of regulator expression (blue red patterns, left) and activity level (orange/purple patterns, right).

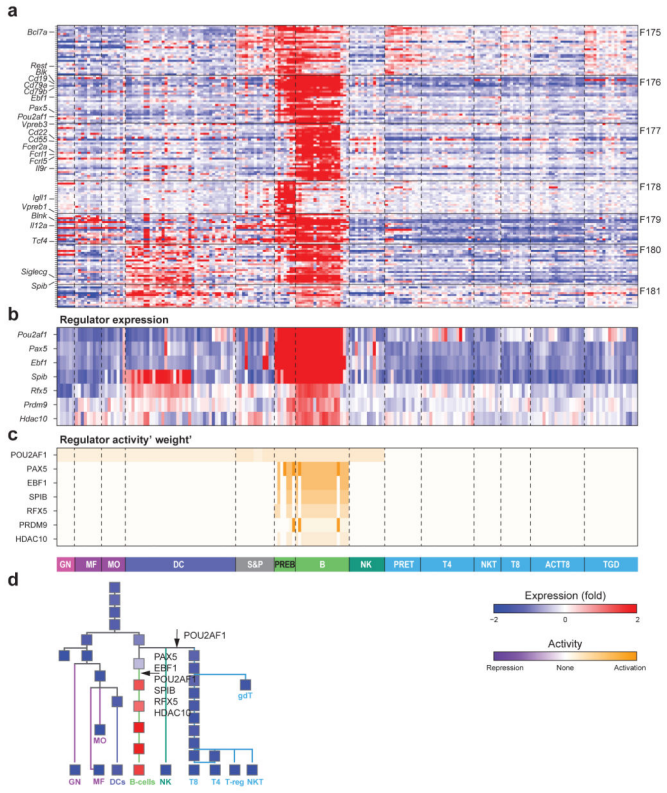

Figure 4.

Ontogenet regulatory model for coarse–grained module C33. (**a**) Module C33. Shown is the mean centered expression (red blue color bar, bottom) of the module's genes (rows) in each cell (column). The major lineages are noted at the bottom, and marked by thin vertical lines. Fine modules F175-F181 nested within C33 are separated by thin horizontal lines and labeled. Example gene names are noted on left. (**b**) Regulators expression. Shown are the mean centered expression levels (red blue color bar, bottom) of the regulators (rows) assigned by Ontogenet to module C33. (**c**) Regulators activity weights. Shown are the activity weights (orange purple color bar, bottom) for each of the Ontogenet assigned regulators from (b) in each cell type. (**d**) Mean-centered mean expression of module C33 is projected onto the hematopoietic tree. Low expression is blue, high expression is red. Selected members are listed below. Selected inferred regulators are marked by arrowhead at the edges in which their activity weight changes. This module contains some typical B cells genes, including *Cd19*, *Blnk*, *Ebf1*, and *Cd79a*.

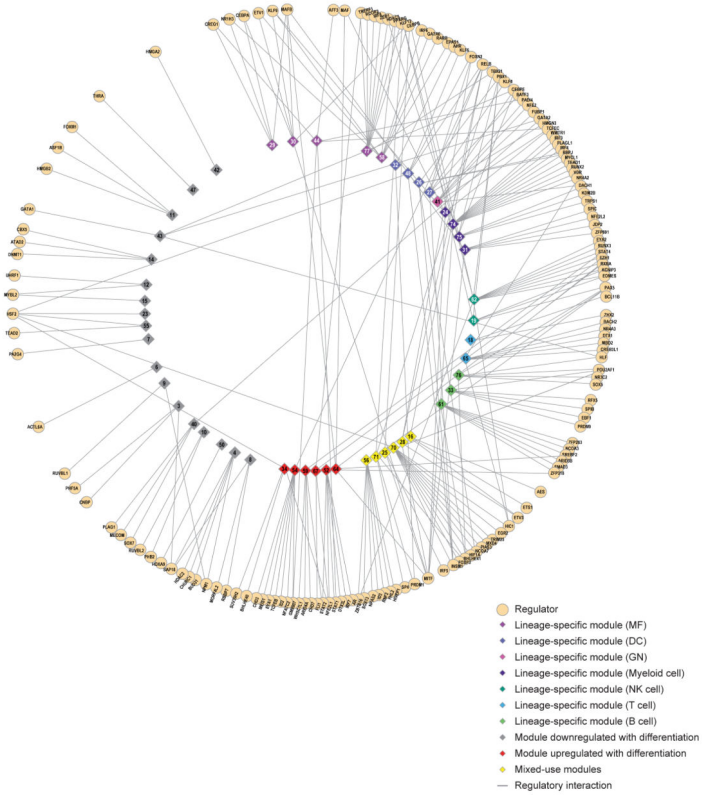

Figure 5.

Ontogenet regulatory model for coarse-grained modules. Shown are lineage specific modules (colored as in Figure 2, except myeloid induced modules, dark purple), pan-differentiation induced (red) and repressed (gray) modules and mixed-used modules (yellow) (all in inner circle), and their Ontogenet assigned regulators (outer circle, cream) with regulatory interactions with maximal effect (absolute activity weight*expression) bigger than 1. An edge connects each regulator to the module(s) it regulates.

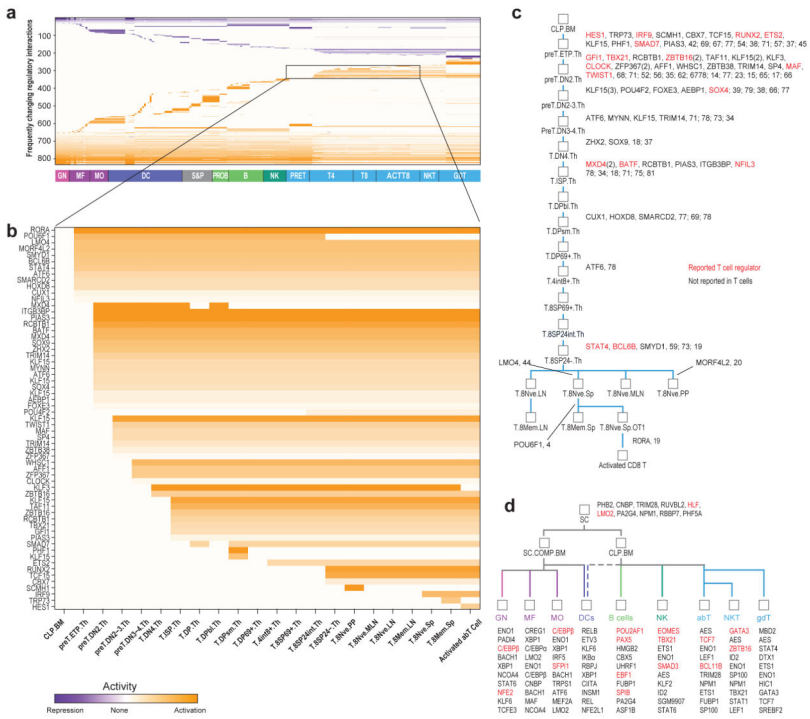

Figure 6.

Changes in activity weights across the hematopoietic lineage tree. (**a**) High changing interactions. Shown are the activity weights in each cell type (column) for every highly changing regulatory interaction between a regulator and a coarse-grained module. Orange: positive (activation) activity weight; Purple: negative (repression) activity weight; White: zero (no regulation). The major lineages are noted by the color bar (bottom). (**b**) High changing interactions in the CD8$^+$ T cell lineage. Shown is a zoom in (from **a**) only for those activating interactions that are recruited within the CD8$^+$ T cell lineage. (**c**) Known and novel regulators recruited in the CD8$^+$ T cell lineage. Shown is the CD8$^+$ T cell lineage branch (squares: cell types; edges: differentiation steps) labeled with the regulators recruited along each differentiation step and their associated modules. Regulators previously reported to have a role in T cells are marked in red. The number of activity weight changes for the regulator on this edge, if more than one, is shown in parentheses. (**d**) Ontogenet-inferred lineage regulators. Shown is a reduced ImmGen tree with the lineage regulators. Regulators previously reported to have a role in that lineage are marked red.

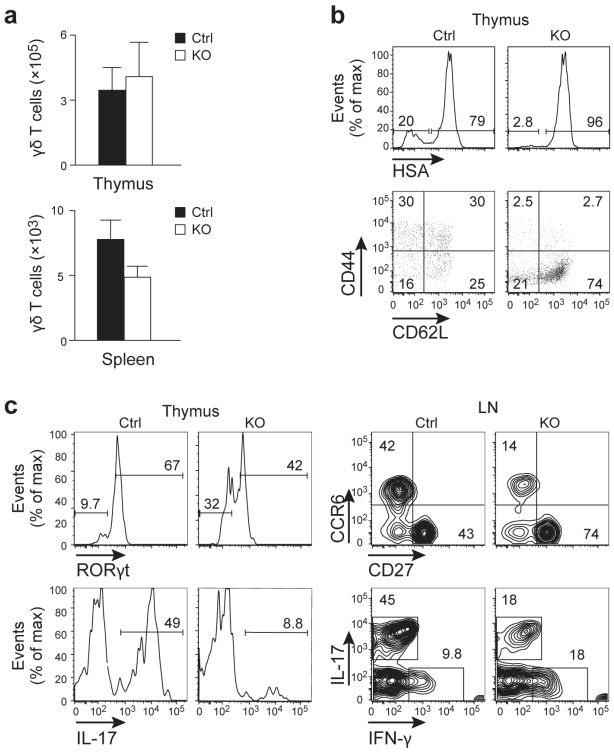

Figure 7.

ETV5 is a γδ T cell regulator. (**a**) Relatively normal total numbers of γδ T cells are generated in ETV5 conditional KO (CKO) mice. In 7 day old neonates, total thymocyte numbers of CKO mice are decreased to ~50% of normal, but the frequency of γδTCR$^+$ thymocytes is increased by ~ 2 fold, resulting in similar numbers of γδ T cells in the thymus and spleen as controls. One representative of three independent litters analyzed with a minimum of two mice/genotype is shown. Control (Ctrl) mice are CD2p-CreTg$^+$*Etv5*$^{+/+}$ littermates. (**b**) Altered maturation of Vγ2$^+$ thymocytes. Top, Decreased numbers of mature HSA (CD24)^{lo} Vγ2$^+$ thymocytes in 7 day old CKO mice. Bottom, Decreased activation of Vγ2$^+$ thymocytes in CKO mice as indicated by the paucity of CD44$^+$ cells. For γδTCR$^+$ thymocytes expressing other Vγ chains, the proportions of mature cells or activated cells in CKO mice were not different from controls. Similar results were observed in mice of different ages. (**c**) Impairment in the capacity to produce IL-17 effector cytokine by Vγ2$^+$ cells. Left panels, decreased expression of IL-17 inducing transcription factor RORγt and corresponding decrease in IL-17 production by mature (CD24^{lo}) Vγ2$^+$ thymocytes in CKO mice. No significant difference in RORγt expression was observed in immature (CD24^{hi}) Vγ2$^+$ thymocytes. Right panels, decreased frequencies and numbers of peripheral CCR6$^+$CD27$^-$IL-17$^+$, and a reciprocal increase in CD27$^+$ IFNγ producing Vγ2$^+$ lymph node (LN) T cells in CKO mice (4 wk old). Data are representative of five experiments.