nature immunology

Identification of transcriptional regulators in the mouse immune system

Vladimir Jojic[1,7], Tal Shay[2,7], Katelyn Sylvia[3], Or Zuk[2], Xin Sun[4], Joonsoo Kang[3], Aviv Regev[2,5,8], Daphne Koller[1,8] & the Immunological Genome Project Consortium[6]

The differentiation of hematopoietic stem cells into cells of the immune system has been studied extensively in mammals, but the transcriptional circuitry that controls 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. To analyze this data set we developed Ontogenet, an algorithm for reconstructing lineage-specific regulation from gene-expression profiles across lineages. Using Ontogenet, we found differentiation stage–specific regulators of mouse hematopoiesis and identified many known hematopoietic regulators and 175 previously unknown candidate regulators, as well as their target genes and the cell types in which they act. Among the previously unknown regulators, we emphasize the role of ETV5 in the differentiation of γδ T cells. As the transcriptional programs of human and mouse cells are highly conserved, it is likely that many lessons learned from the mouse model apply to humans.

The Immunological Genome Project (ImmGen) is a consortium of immunologists and computational biologists who aim, through the use of shared and rigorously controlled data-generation pipelines, to exhaustively chart gene-expression profiles and their underlying regulatory networks in the mouse immune system[1]. In this context, we provide the first comprehensive analysis of the ImmGen compendium and use a new computational algorithm to reconstruct a modular model of the regulatory program of mouse hematopoiesis. Understanding the regulatory mechanisms that underlie the differentiation of cells of the immune system has important implications for the study of development and for understanding the basis of human immunological disorders and hematological malignancies. Most studies of hematopoiesis view differentiation as a process controlled by relatively few 'master' transcription factors that are expressed in specific lineages and act to set and reinforce distinct cell states[2]. However, analysis of gene expression in 38 cell types in human hematopoiesis[3] has suggested a more complex organization that involves a larger number of transcription factors that control combinations of modules of coexpressed genes and are arranged in densely interconnected circuits. However, that human study was restricted to human cells that could be obtained in sufficient quantities from peripheral or cord blood and thus could not access many cell populations of the immune system.

The 246 cell types of the mouse immune system in the 816 arrays of the ImmGen compendium offer an unprecedented opportunity for studying the regulatory organization of hematopoiesis in the context of a rich and diverse lineage tree. Because the transcriptional programs of human and mouse cells are highly conserved[4], many lessons learned from the mouse model will probably be applicable to humans. Two key approaches for the identification of regulatory networks[5] are physical models based on the association of a transcription factor or a *cis*-regulatory element with a target's promoter (for example, from chromatin immunoprecipitation (ChIP) followed by deep sequencing) and observational models with which regulation can be inferred from a significant correlation between the abundance or activity of a transcription factor (as protein or mRNA) and that of its presumed target. In both cases, analysis of the relationship between a putative regulator and a module of coregulated targets enhances robustness and biological interpretability[6,7]. Physical data provide direct evidence of biochemical interactions but do not necessarily indicate function[8] and are challenging to collect[9], whereas mRNA profiles are highly accessible but provide only correlative evidence. As physical and observational models are complementary, using both[3] can enhance confidence[5–7] and expand the scope of discovery.

Analysis of cells organized in a known lineage, as in hematopoiesis, offers unique opportunities that have not been leveraged before. In particular, published models[3] have not explicitly considered the fact that cells that are more closely related (according to the known lineage tree) probably share many of their regulatory mechanisms and that regulatory relationships that exist in one sublineage may not be active in another. Incorporating such information may help in the identification of true regulators of hematopoiesis.

Here we used those insights to develop a new computational method, Ontogenet, and to apply it to the ImmGen compendium

[1]Computer Science Department, Stanford University, Stanford, California, USA. [2]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [3]Department of Pathology, University of Massachusetts Medical School, Worcester, Massachusetts USA. [4]Laboratory of Genetics, University of Wisconsin-Madison, Madison, Wisconsin, USA. [5]Howard Hughes Medical Institute, Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [6]A full list of members and affiliations appears at the end of the paper. [7]These authors contributed equally to this work. [8]These authors jointly directed this work. Correspondence should be addressed to A.R. (aregev@broadinstitute.org) or D.K. (koller@cs.stanford.edu).

Received 30 January; accepted 13 March; published online 28 April 2013; doi:10.1038/ni.2587

to build an observational model associating 578 candidate regulators with modules of coexpressed genes. We defined modules at two different granularities, with 81 larger coarse-grained modules, some of which we further refined into smaller modules with more coherent expression; this resulted in 334 fine-grained modules. The model identified many of the already known hematopoietic regulators, was supported through the use of a complementary physical model and proposed dozens of previously unknown candidate regulators. Our model provides a rich resource of testable hypotheses for experimental studies, and the Ontogenet algorithm can be used to delineate regulation in the context of any cell lineage.

RESULTS

Transcriptional compendium of the mouse immune system

The ImmGen consortium data set[1] (April 2012 release) consists of 816 expression profiles from 246 cell types of the mouse immune system (**Fig. 1** and **Supplementary Table 1**). The cell types span all major hematopoietic lineages, including stem and progenitor cells, granulocytes, monocytes, macrophages, dendritic cells (DCs), natural killer (NK) cells, B cells and T cells. The T cells include many types of αβ T cells, regulatory T cells (T_{reg} cells), natural killer T cells (NKT cells) 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 were largely consistent with the known lineage tree (**Fig. 2**). In general, the closer two cell populations were in the lineage tree, the more similar their expression profiles were (Pearson $r = -0.71$; **Supplementary Fig. 1**). For myeloid cells, profiles were similar overall, with granulocytes being the least variable, DCs being the most variable (consistent with DC samples' being obtained from diverse tissues and their known inherent diversity[10]) and all myeloid cells being weakly similar to stromal cells. Conversely, lymphocytes had larger differences between lineages. NK cells, although tightly correlated, did show weaker similarity to T cells, especially $CD8^+$ T cells and NKT cells. T cells were very heterogeneous, which partly reflected the finer sampling for this lineage. Stem cells were most similar to early myeloid and lymphoid progenitors (S&P group, **Fig. 2**), followed by pre-B cells and pre-T cells, consistent with a gradual loss of differentiation potential. As a resource for studying each lineage, we used one-way analysis of variance to define characteristic signatures of over- and underexpressed genes for each of the main eleven lineages compared with the expression of those genes in 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 coexpressed genes at two granularities (**Supplementary Fig. 2a**,**b**). We first constructed 81 coarse-grained modules (C1–C81;

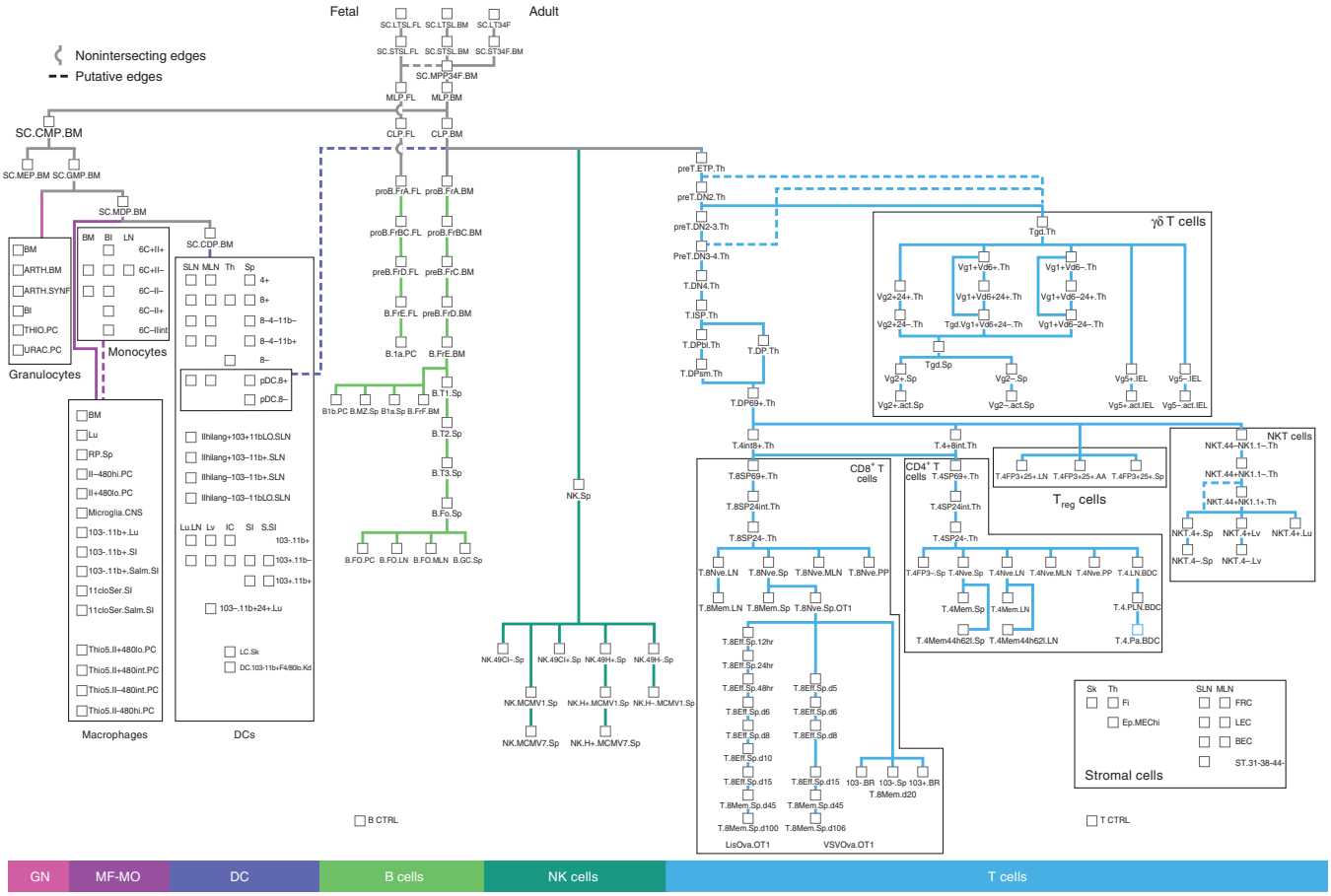

Figure 1 Mouse cell populations in the ImmGen compendium. Lineage tree of the hematopoietic mouse cell types profiled by the ImmGen Consortium (nomenclature and markers for sorting, **Supplementary Table 1**). Some samples of stem cells, progenitor cells and B cells were obtained from adult and fetal liver. Stromal cells (box, bottom right) were also measured as part of ImmGen but are not part of the lineage tree. Color in bar beneath matches color of branches in tree above. GN, granulocyte; MF-MO, macrophages-monocyte. Adapted from ref. 4.

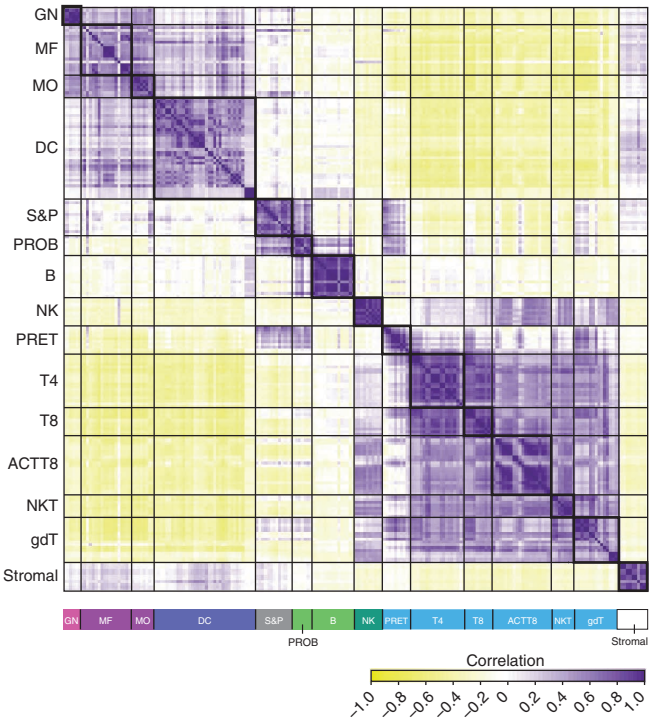

Figure 2 Related cells have very similar expression profiles. Pearson correlation coefficients (purple, positive; yellow, negative; white, none) for each pair of profiled cell types, calculated for the 1,000 genes (of the 8,431 unique expressed genes) with the highest s.d. value of all samples. Samples are sorted by breadth-first search on the tree (**Fig. 1**), with stromal cells at the lower or right end. Black vertical and horizontal lines delineate major lineages according to labels along left margin and beneath (color in bar beneath matches colors in **Fig. 1**). S&P, stem and progenitor cells; PROB, pre-B cells and pro-B cells; T4, $CD4^+$ T cells; T8, $CD8^+$ T cells; ACTT8, activated $CD8^+$ T cells; gdT, γδ T cells.

Supplementary Fig. 2c–h and **Supplementary Table 3**) and then further identified for each coarse-grained module a set of nested fine modules (**Supplementary Fig. 2a**), which resulted in 334 fine modules spanning 7,965 genes (F1–F334; **Supplementary Table 4**). Coarse modules helped us capture the mechanisms that coregulate a larger set of genes in one lineage, whereas fine modules may help in the identification of distinct regulatory mechanisms that control only a smaller subset of these genes in the other lineage(s). Many of the modules showed enrichment for coherent functional annotations, *cis*-regulatory elements (**Supplementary Table 5**) and binding of transcription factors (**Supplementary Table 6** and **Supplementary Note 1**), including binding sites for factors known to act as regulators in 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 at the ImmGen portal (http://www.immgen.org/ModsRegs/modules.html).

Most coarse-grained modules (48 of 81 modules; 4,478 of 7,965 genes) showed either lineage-specific induction (**Supplementary Figs. 2c** and **3**) or 'pan-differentiation' regulation (**Supplementary Figs. 2de**, **4** and **5**). In addition, 6 modules were 'mixed-use' across lineages (**Supplementary Figs. 2f** and **6**), 8 were stromal specific (**Supplementary Fig. 2g**) and 19 had expression patterns that did not fall into those categories (**Supplementary Figs. 2h** and **7**). Lineage-specific repression was rare (only in C53 (B cells) and C17 (stromal cells)).

Ontogenet: reconstructing lineage-sensitive regulation

We next developed a new algorithm, Ontogenet, to delineate the regulatory circuits that drive hematopoietic cell differentiation. Ontogenet aims to fulfill the following 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 of those factors may change in different cell types (for example, 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 that regulate a module are more similar in cells that are close to each other in the lineage tree (for example, from the same sublineage) than in 'distant' cells (for example, from two different sublineages), in accordance with the greater similarity in expression profiles of closer cell types (**Supplementary Fig. 1**); and criterion 4, master regulators of a lineage (for example, GATA-3 for T cells) are active across the sublineages, but the subtypes can also have additional, more specific regulators (for example, Foxp3 for T_{reg} cells). The former should be captured as shared regulators of a coarse module and its nested fine modules, whereas the latter regulate only particular fine modules.

Ontogenet receives as input the gene-expression module, the lineage tree and the expression profiles of a predesignated set of 'candidate regulators' (transcription factors, chromatin regulators and so on). It then associates each module with a combination of regulators (criterion 1 above), whereby each regulator is assigned an 'activity weight' for each cell type that indicates 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 abundance. Following the approach in the Lirnet method for regulatory-network reconstruction[6], the activity-weighted expression of the regulators is combined in a linear model to generate a prediction of a module's 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 is 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 genes in the same module are regulated in the same way. This is essential for statistical robustness, although it comes at the 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.

Although Ontogenet reconstructs a potentially different regulatory program for each cell type, as reflected by the cell-specific activity weights of each regulator, it is geared toward maintaining the same activity 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). Collectively, we use an optimization approach that constructs an ensemble of regulatory programs that try to achieve the following 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 regulators similar to those of the coarse modules to which they belong.

Notably, the approach used before to identify combinations of regulators (for example, linear regression regularized with the Elastic Net

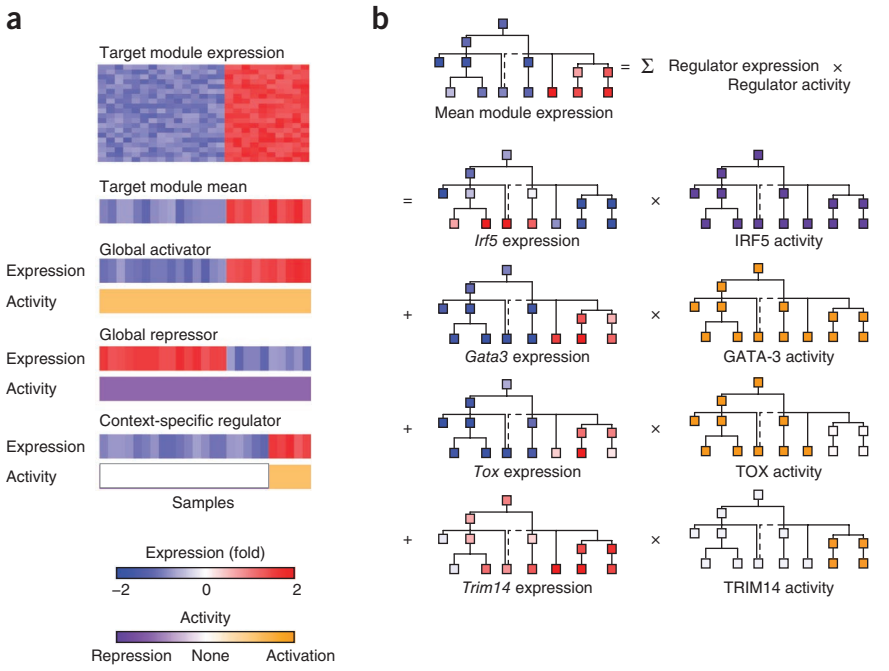

Figure 3 Overview of Ontogenet. (**a**) Use of the regulator expression profile (blue-red; key below) and activity profile (orange-purple; key below) to demonstrate how a type of regulator (bottom) can 'explain' the expression of a module (top): a regulator may have a uniformly positive activity weight across the lineage (constitutive activator; top), a uniformly negative activity weight (constitutive repressor; middle) or variable activity weights (context-specific regulator; bottom). (**b**) Mean expression of a module (top) calculated as a linear combination of regulator expression (blue-red; left) and activity (orange-purple; right).

penalty[6,11]) assumed that regulatory activity (and hence activity weight) is the same across all cell types. Thus, if a regulator was expressed similarly in two different cells, it was deemed to be 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 between related cell types in the lineage in terms of gene regulation. 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 modules and 334 fine modules, a lineage tree consisting of 195 cell types and 580 candidate regulators. The Ontogenet model identified 1,417 regulatory relations (1,091 activating, 317 repressing and 9 mixed) between 81 coarse-grained modules and 480 unique regulators (**Fig. 4**, **Supplementary Fig. 8** and **Supplementary Table 5**). 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 at predicting new and previously unknown 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** and **Supplementary Note 3**).

In most cases (59%), a regulator's activity weights varied in different cell types ('frequently changing'), reflective of context-specific regulation (**Supplementary Fig. 10**). When we pruned regulatory interactions whose maximal effect (defined as the product of activity weight and expression) was low, we obtained a sparser network, in which 'pan-differentiation' and lineage-specific modules were controlled mostly by distinct regulators (**Fig. 5**), whereas mixed-use modules shared 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 activity (that is, a regulator's activity weights frequently changed in some modules and remained constant in others), probably reflective of both the greater number of interactions and the finer regulatory program (**Supplementary Fig. 10** and **Supplementary Table 7**). This rich regulatory model for differentiation of the mouse immune system identified many known regulatory interactions and suggested new regulatory interactions in specific immunological contexts.

Ontogenet prediction of known regulatory interactions

Many of the regulatory interactions identified by Ontogenet were already known, which supported the accuracy of our model. For example, among individual regulators, PU.1 (encoded by *Sfpi1*) was selected as a regulator of the myeloid and B cell 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 the myeloid-specific modules C25 and C30 and many myeloid fine modules; MafB (encoded by *Mafb*) regulates the macrophage-specific modules C29, F128 and F131; STAT1 regulates the interferon-response module C52; T-bet (encoded by *Tbx21*) regulates the NK cell module C19 and NKT cell module F288; and CIITA (encoded by *Ciita*) regulates the antigen-presenting cell module F136.

Furthermore, the combination of regulators associated with a single module was 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 Spi-B (**Fig. 4**); the T cell module C18 (**Supplementary Fig. 8**) is regulated by the known T cell regulators Bcl-11B, GATA-3, Lef1, TOX and TCF7; the γδ T cell module C56 is regulated by the known γδ T cell regulators PLZF (ZBTB16), Sox13 and Id3, all also involved in NKT cell development and function; the NKT module F188 is regulated by GATA-3, T-bet and PLZF; and fine modules F150 and F152, in which the expression of their member genes by $CD8^+$ DCs is higher than that of $CD4^+$ DCs, 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 were 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**), which supported the idea of 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 were supported by enrichment for *cis*-regulatory motifs ($P = 2.6 \times 10^{-5}$ (hypergeometric test for two groups) and $P < 1 \times 10^{-5}$ (permutation test)), such as

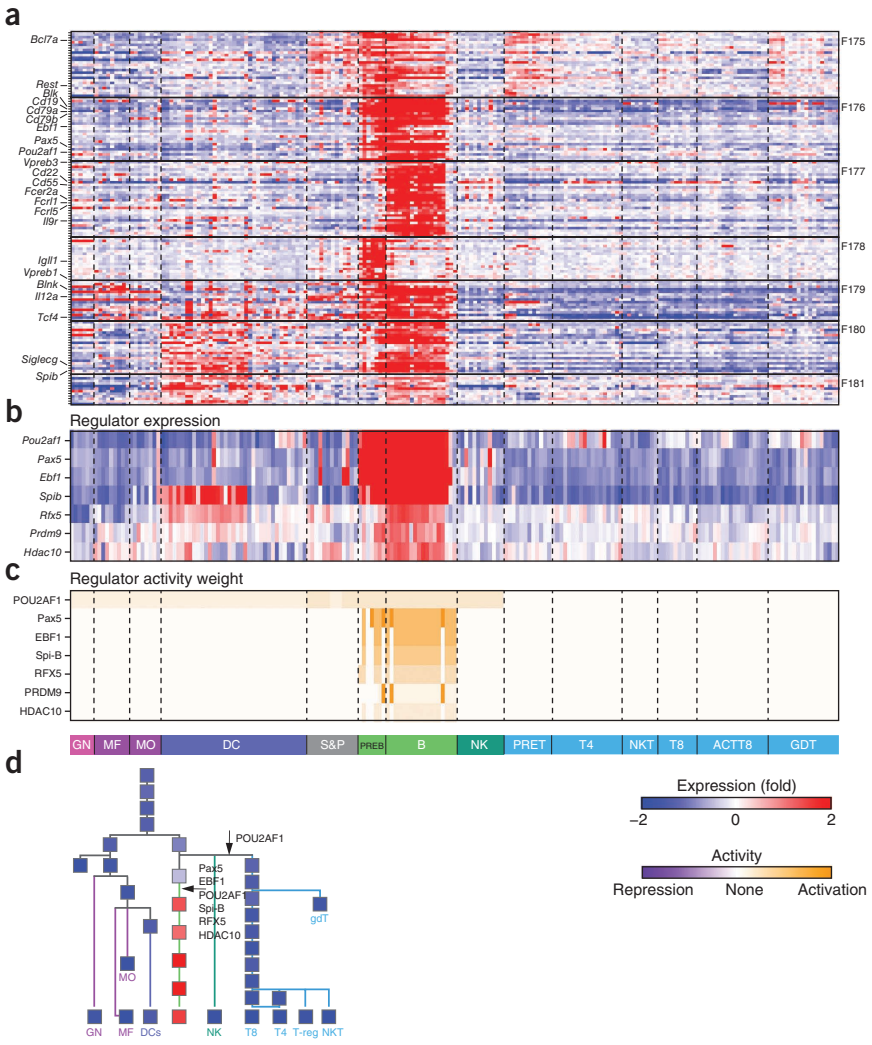

Figure 4 Ontogenet regulatory model for coarse–grained module C33. (**a**) Module C33: mean-centered expression (red-blue; key, bottom right) of the module's genes (rows) in each cell (column); major lineages are delineated by dashed vertical lines (which correspond to color bar beneath **c**; matches colors in **Fig. 1**); fine modules F175–F181 (right margin) nested within C33 are delineated by thin horizontal lines; left margin, examples of genes in module. This module contains some typical B cell genes, including *Cd19*, *Blnk*, *Ebf1* and *Cd79a*. (**b**) Regulator expression, presented as mean-centered expression (red-blue color bar, below **c**) of the regulators (rows) assigned by Ontogenet to module C33. (**c**) Regulator activity weight (orange-purple; key, bottom right) assigned by Ontogenet for each of the regulators from **b** in each cell type. (**d**) Projection of mean-centered mean expression (blue, low; red, high) of module C33 onto the hematopoietic tree (below, differentiated populations on that tree); arrowheads indicate 'edges' (differentiation steps) at which the activity weight of selected inferred regulators (labeled in diagram) changes.

the GATA-2 motif in the hematopoietic stem cell (HSC) module C40, and the PU.1 (SFPI1) motif in myeloid cell module C25. The ChIP profiles supported the prediction of 21 regulator–coarse module associations ($P = 2.2 \times 10^{-5}$ (hypergeometric test for two groups) and $P < 1 \times 10^{-5}$ (permutation test)), such as the binding of C/EBPα and C/EBPβ in the myeloid cell module C24 and the binding of EBF1 in the B cell module C33.

Although those overlaps were statistically significant, they nevertheless also indicated that the predictions of most regulatory interactions were not supported by enrichment for known *cis*-regulatory motifs or available transcription factor–binding data, and vice versa. There are three reasons for this. First, assigning scores for binding sites and their enrichment is a process that is highly prone to false-negative results; this is particularly likely to occur for 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) or ChIP binding data in any cell type (90% of regulators; 497 of 554). Such regulators can be nominated only by an expression-based method, such as Ontogenet, and should not be considered false-positive results of our method. Third, in many cases in which we do find enrichment for a *cis*-regulatory element or binding profile for (for example) transcription factor A in module B (300 of 551 *cis*-regulatory interactions (54%); 52 of 90 ChIP-based interactions (57%)), the transcription factor (A) and its target module (B) show little or no correlation in expression (absolute Pearson $r < 0.5$). In some cases, this is due to a factor that is not itself transcriptionally regulated (a real 'false-negative' result of Ontogenet), but in many other cases the factor probably controls these targets in another cell type not measured in our study (and hence is not in fact a false-negative result of Ontogenet).

A few known regulators of differentiation of the immune system[13] were not identified by the model for various reasons. Tal-1 and BMI1 did not meet the initial filtering criteria, as they were expressed only 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 was highest in pre-T cells and was 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 was not specifically expressed in T cells and had low expression in general, possibly because of a bad probe set. XBP1 was not identified as a B cell regulator because it had relatively low expression in B cells in our arrays and had higher expression in myeloid cells.

The reidentification of known regulators lends support to the many previously unknown regulatory interactions in the model. Of the 475 regulators that Ontogenet associated with lineage-specific modules or 'pan-differentiation' modules, at least 175 (37%) were completely unknown in this context. Among those, for example, KLF12 was predicted to be a regulator of the NK cell module C19 but was not associated before with the regulation of NK cells. GATA-6 was predicted to be a regulator of the macrophage-specific modules C31, C50 and C58 but was not associated before with macrophages. That is in agreement with the much lower number of granulocyte-macrophage colonies generated by embryoid bodies of GATA-6-deficient mice[14]. Finally, ETV5 was predicted by the model to be a regulator of the γδ T cell modules F287 and F289, a previously unknown role discussed below.

Context-specific regulation underlies mixed-use modules

Context-specific regulation, in which the same set of genes is regulated by one set of regulators in the context of one lineage and by

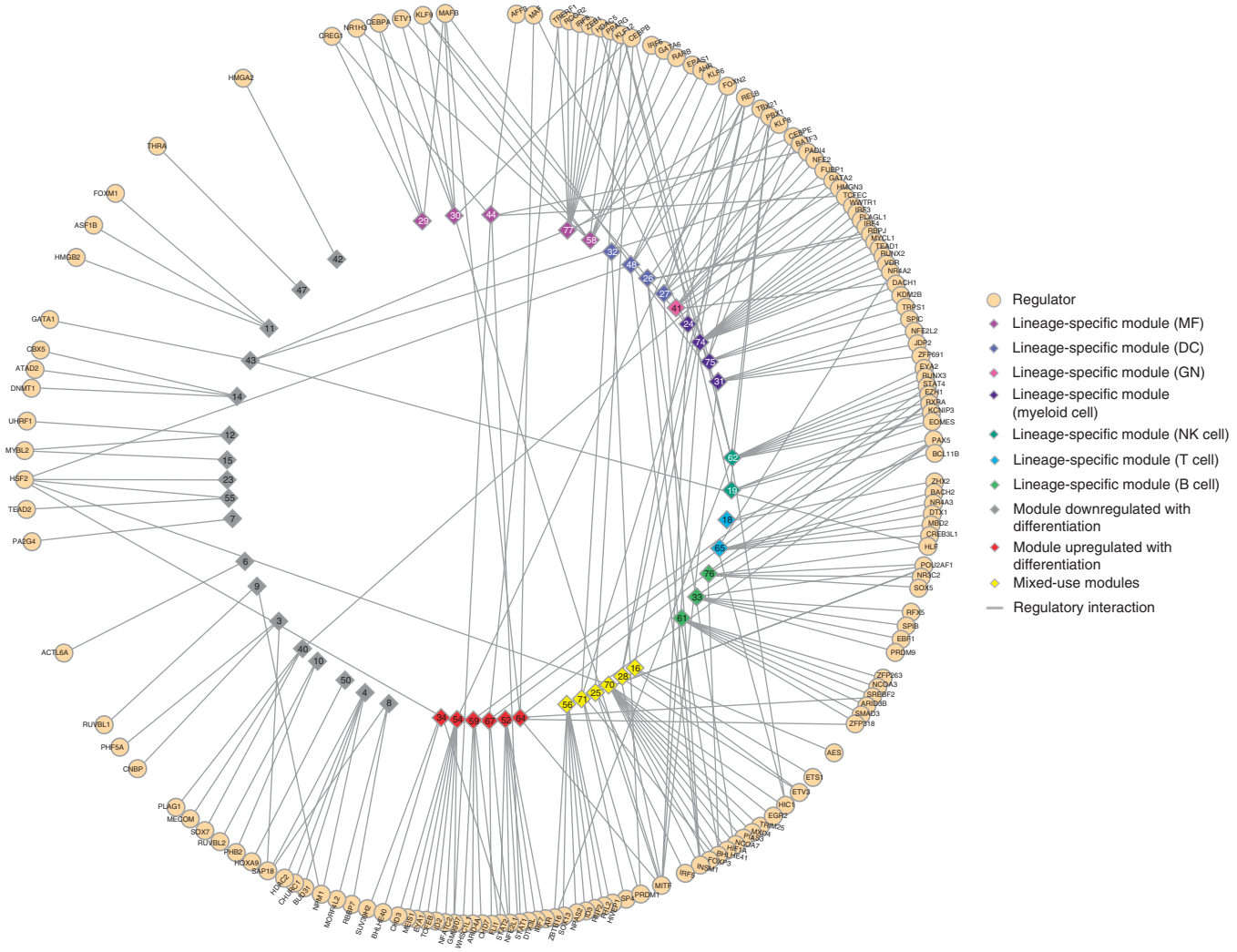

Figure 5 Ontogenet regulatory model for coarse-grained modules. Lineage-specific modules (inner circle; colors as in **Fig. 2**, except myeloid induced modules (dark purple)), 'pan-differentiation' induced (red) and repressed (gray) modules and mixed-use modules (yellow) and their Ontogenet assigned regulators (outer circle; cream) with regulatory interactions with a maximal effect (absolute activity weight × expression) >1. Lines (regulatory interactions) connect each regulator to the module(s) it regulates.

another set of regulators in the context of another lineage, has been reported in selected cases, such as the regulation of *Rag2* by GATA-3 in T cells and by Pax5 in B cells[15]. The ability of Ontogenet to identify different regulatory programs for the same module in different parts of the lineage tree can help delineate the regulatory mechanisms that underlie 'mixed-use' modules expressed in more than one lineage. For example, module C70 is induced both in T_{reg} cells 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, the fine-grained module F300 is independently induced in both mature B cells and T cells. Although some of its regulators are themselves 'mixed-use' in both lineages, others are B cell specific (ZFP318, RFX5 and CIITA) or T cell specific (EGR2).

Regulatory recruitment and 'rewiring' during differentiation

Most regulatory relations identified by Ontogenet were dynamic, as reflected by the change in their associated activity weights during differentiation. This change provided 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 changed (increased or decreased) between that cell type and its immediate progenitor (**Supplementary Table 9**), as well as the unique regulators and modules involved in those interactions. In this way, we identified modules and regulators that were recruited and strengthened (activity weight greater than that of its progenitor) or were disposed of and weakened (activity weight lower than that of its 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 common lymphoid progenitors, 61 regulatory interactions were recruited, involving 34 modules and 49 regulators, only 15 of which have been associated before with T cell differentiation. In particular, for the differentiation step from double-negative ($CD4^-CD8^-$) stage 4 T cell to immature single-positive ($CD4^+$ or $CD8^+$) T cell, Ontogenet independently identified the previously reported involvement of MXD4, Batf and NFIL3 and newly identified the involvement of RCBTB1, PIAS3 and ITGB3BP (**Fig. 6b**,**c**).

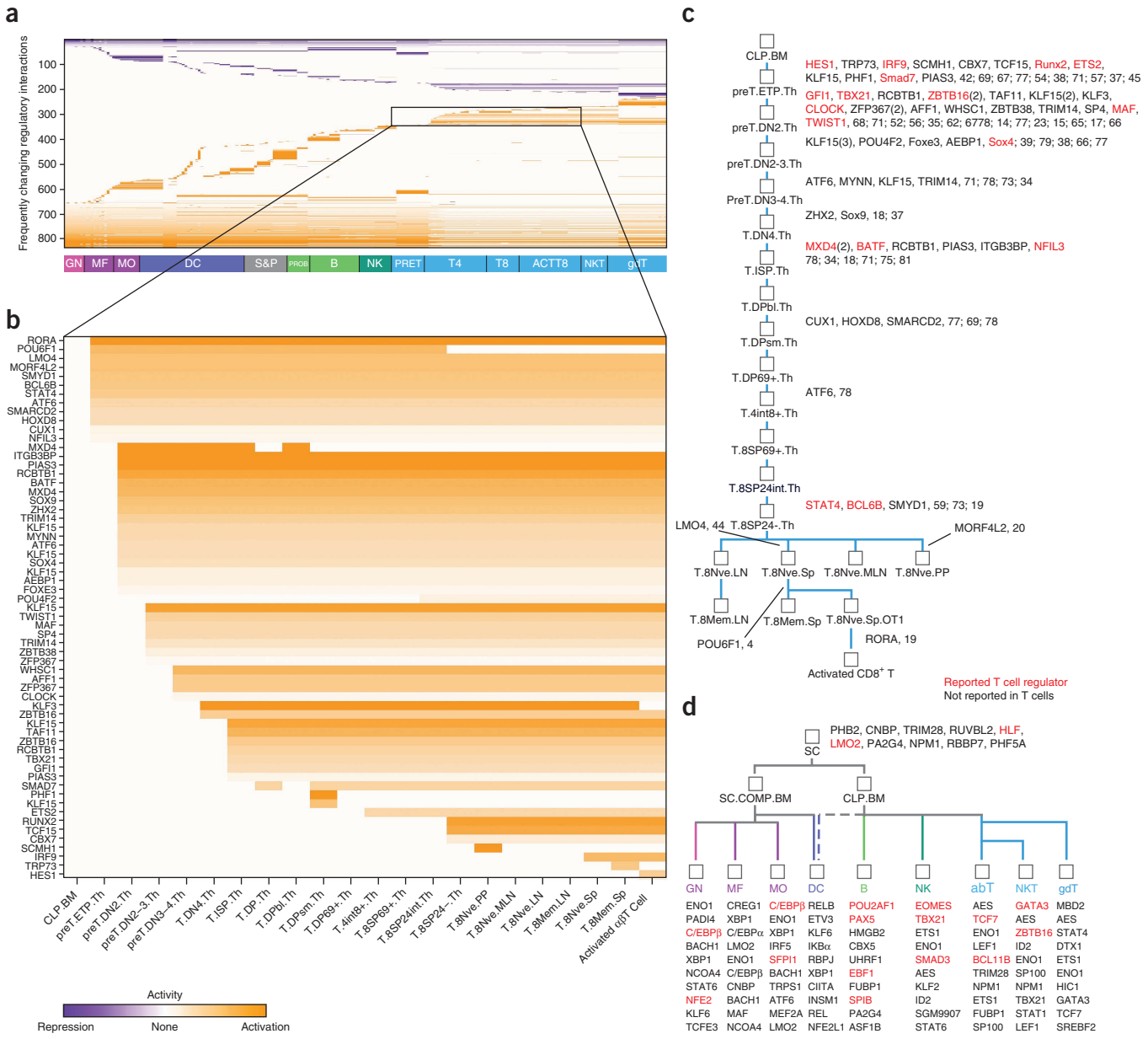

Figure 6 Changes in activity weights across the hematopoietic lineage tree. (**a**) Activity weights in each cell type (columns; correspond to color bar below, which matches colors in **Fig. 1**) for every 'frequently changing' regulatory interaction between a regulator and a coarse-grained module: orange, positive (activation) activity weight; purple, negative (repression) activity weight; white, 0 (no regulation). (**b**) Enlargement of boxed area in **a** of 'frequently changing' interactions for only those activating interactions recruited in the CD8$^+$ T cell lineage. (**c**) Known and previously unknown regulators recruited in the CD8$^+$ T cell lineage, including the CD8$^+$ T cell lineage branch (squares, cell types; lines ('edges'), differentiation steps); right, regulators recruited along each differentiation step and their associated modules (red, regulators reported to have a role in T cells; black, not reported in T cells). Numbers in parentheses indicate activity weight changes for the regulator on this edge, if >1. (**d**) Simplified ImmGen tree (cell type colors correspond to those in **Fig. 1**) with Ontogenet-inferred lineage regulators (red and black as in **c**).

In another example, during the differentiation step that leads to NK cells, the NK cell module C19 was assigned the known NK cell regulators Eomes and T-bet as activators. Both Eomes and T-bet were also recruited as repressors at this differentiation step in other modules. The differentiation step that leads to T_{reg} cells recruited the T_{reg} cell module C70 and its known regulators Foxp3 and CREM (which has been proposed as a T_{reg} cell regulator[16]). Notably, because HSCs have no parent in our model, regulators active in HSCs will be noted only when they are no longer used at later points (for example, HOXA7 and HOXA9 were no longer used as activators at the multilymphoid progenitor stage). The first differentiation step with activator recruitment is the step that leads to multilymphoid progenitors, at which MEIS1 is recruited to module C42. MEIS1 is later no longer used by C42 in T cells, in agreement with the reported methylation and silencing of the gene encoding MEIS1 during differentiation toward 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

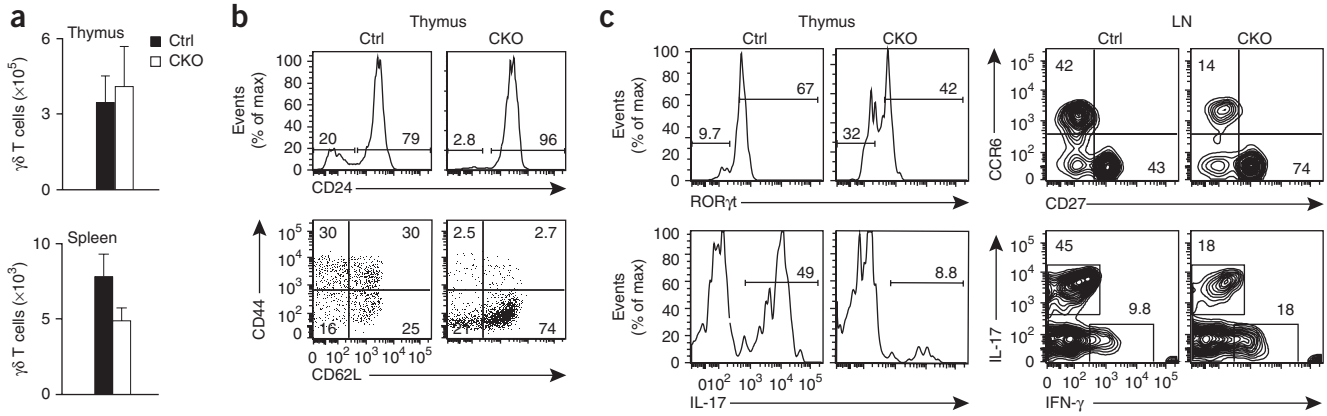

Figure 7 ETV5 is a regulator of γδ T cells. (**a**) Total γδ T cells in the thymus and spleen of mice with T cell–specific ETV5 deficiency (CKO) and their CD2p-CreTg^{+}*Etv5*$^{+/+}$ littermates (control (Ctrl)). (**b**) Expression of CD24 (top) and of CD44 and CD62L (bottom) by $V_{\gamma}2^{+}$ thymocytes from 7-day-old mice as in **a**. Numbers above bracketed lines (top) indicate percent CD24^{lo} (mature) thymocytes (left) or CD24^{hi} (immature) thymocytes; numbers in quadrants (bottom) indicate percent cells in each. Similar results were obtained for mice of different ages. (**c**) Expression of RORγt and production of IL-17 (left) by mature (CD24^{lo}) $V_{\gamma}2^{+}$ thymocytes of mice as in **a**; numbers above bracketed lines indicate percent RORγt^{-} cells (top; left number) or RORγt^{+} cells (top; right number), or IL-17^{+} cells (bottom). Right, expression of CCR6 and CD27 (top) and of IL-17 and interferon-γ (IFN-γ; bottom) by $V_{\gamma}2^{+}$ T cells from the lymph nodes (LN) of 4-week-old mice as in **a**; numbers in or adjacent to outlined areas indicate percent cells in each. Data are representative of three experiments with one to three independent litters with a minimum of two mice per genotype (**a**; error bars, s.e.m.), five experiments (**b**) or three experiments (**c**).

and repressors on the basis of the entire model (**Fig. 6d** and **Supplementary Table 10**). In this way we correctly captured many known regulators of each lineage among the top-ranked activators. For example, our model associated c-Myc, N-Myc, GATA-2 and MEIS1 with stem and progenitor cells; Bcl-11B, TCF7 and GATA-3 with αβ T cells; POU2AF1, Pax5, EBF1 and Spi-B with B cells; Eomes, T-bet and Smad3 with NK cells; and GATA-3 and PLZF with NKT cells. In addition, the model made many predictions of lineage regulators not previously associated with those lineages, such as the following: in stem and progenitor cells, HLF; in granulocytes, DACH1 (reported to regulate cell-cycle progression in myeloid cells[18]), Bach1 and NFE2; in macrophages, CREG1; in DCs, ATF6, ETV3, SKIL, NR4A2 and NR4A3 (shown to be induced in viral infected DCs[19,20]); in monocytes, POU2F2 (Oct2; reported to be upregulated during macrophage differentiation[21]) and KLF13 (a regulator of B cells and T cells[22] with higher expression in monocytes); in B cells, ZFP318; and in NK cells, ELF4 (Gm9907; shown to control the proliferation and homing of CD8^{+} T cells[23]). Notably, although this 'pan-model' analysis is useful, it can deemphasize 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 presented above (**Fig. 6**).

Finally, by counting the changes in activity weight that occur (across all regulators and modules) at each differentiation step ('edge'), we can identify those differentiation points at which regulation is 'rewired' most substantially (**Supplementary Fig. 11**). For example, 19 regulators were recruited to coarse modules (that is, their activity weight increased from 0) at the early thymocyte progenitor stage, and 28 regulators were recruited to coarse modules at the γδ T cell stage, including the known T cell regulator GATA-3 and the known γδ T cell regulators Id3 and Sox13 (**Supplementary Fig. 11a**). At the common lymphoid progenitor stage, four regulators were disposed of (that is, their activity weight diminished to 0) by coarse modules, including the HSC regulators HOXA7, HOXA9 and HOXB3. Eighteen regulators were disposed of at the double-negative-2 T cell precursor stage, including GATA-1, c-Myc and N-Myc (**Supplementary Fig. 11b**). Overall, 'rewiring' was more prominent at higher levels in the lineage than at lower (more-differentiated) levels, although this may have been partly due to the diminished power to detect changes in cell types with no other cells differentiating from them (terminally differentiated; also called 'leaves in the tree'). The individual differentiation steps with the largest number of activity weight changes were those in small-intestine DCs, thymus γδ T cells, liver and lung DCs and double-negative-2 T cell precursor stage, which suggests substantial regulatory 'rewiring' in these cells, possibly due to tissue-specific effects. The regulatory model for fine modules identified a larger number of regulatory changes (a change in activity weight for 82% of the differentiation steps, compared with 65% for the coarse-grained module model), in particular in differentiation steps leading to 'leaves' (terminally differentiated cells; 67% versus 48%). Thus, the fine-grained modules help to identify more cell type–specific regulation.

ETV5 regulates γδ T cell differentiation

To test one of the model's predictions *in vivo*, we centered on regulatory activators of lineage-specific modules with no known function in that lineage. A practical criterion was that the gene could be manipulated *in vivo* in a cell type–restricted manner. We focused on the Ets family member ETV5 and its predicted role as a regulator of the differentiation of γδ T cells in modules F287 and F289, as its expression is highly restricted to the γδ T cell lineage. Although the model assigned 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, which raised 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 linked to γδ T cell development thus far.

To assess the regulatory role of ETV5 in γδ T cells, we analyzed γδ T cell development and function in mice lacking ETV5 specifically in T cells (CD2p-*Cre*Tg^{+}*Etv5*$^{fl/fl}$ mice). As thymocytes that express the γδ T cell antigen receptor transit from immature cells with high expression of the cell surface marker CD24 (CD24^{hi}) to mature CD24^{lo} cells, they acquire effector functions[24]. ETV5 has its highest expression in γδ thymocytes expressing γ-chain variable region 2

($V_\gamma2$) of the T cell antigen receptor, which constitute nearly half of all γδ T cells in postnatal mice. Most $V_\gamma2^+$ cells differentiate into interleukin 17 (IL-17)-producing γδ effector cells in the thymus[24]. Thus, one prediction of the model was that the intrathymic development of IL-17-producing γδ effector cells would be particularly impaired in the absence of ETV5. In mice with conditional T cell–specific deficiency in ETV5, the overall number of γδ T cells generated was similar to that of control mice (their CD2p-CreTg$^+$*Etv5*$^{+/+}$ littermates): in 7-day-old neonates, total number of thymocytes in mice with T cell–specific ETV5 deficiency was ~50% of normal, but the frequency of thymocytes that expressed the γδ T cell antigen receptor was about twofold higher, which resulted in an abundance of γδ T cells in the thymus and spleen similar to that in control mice (**Fig. 7a**). However, there was specific loss of mature $V_\gamma2^+$ thymocytes in mice with T cell–specific ETV5 deficiency (**Fig. 7b**, top). This may have been due to inefficient activation, as indicated by the lower expression of CD44 (the nominal marker of lymphocyte activation) on $V_\gamma2^+$ thymocytes from mice with T cell–specific ETV5 deficiency and the correspondingly higher expression of CD62L (a marker of the naive state) on those cells (**Fig. 7b**, bottom). For γδ thymocytes that expressed other V_γ chains, the proportion of mature cells or activated cells in mice with T cell–specific ETV5 deficiency was not different from that of controls. Critically, the residual mature thymocytes in mice with T cell–specific ETV5 deficiency were impaired in the generation of IL-17-producing γδ effector cells (**Fig. 7c**). Mature $V_\gamma2^+$ thymocytes from ETV5-deficient mice had lower expression of the transcription factor RORγt (which induces *Il17* transcription), and both thymic and peripheral γδ T cells were impaired in the generation of CCR6$^+$CD27$^-$ IL-17-producing γδ effector cells (**Fig. 7c**). These results supported the prediction of our model and demonstrated that ETV5 was essential for proper intrathymic 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 of the ImmGen portal (http://www.immgen.org/ModsRegs/modules.html) is the gateway to the Ontogenet regulatory model of the ImmGen. It allows the user to browse coarse-grained or fine-grained modules by their number, their pattern of expression, a gene they contain, a regulator 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 heat map), the activity weights of each regulator in each cell, and the module's mean expression projected on the lineage tree (as in **Fig. 4a**). 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 for downloading 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 data set provides the most detailed and comprehensive view of the transcriptional activity of any mammalian immune system and (arguably) of any developmental cell–differentiation process. We have used those data to analyze the regulatory circuits underlying such processes, from global profiles to modules to the transcription factors that control them. The unique features of Ontogenet have allowed us to identify regulatory programs active at specific differentiation stages and to follow them as they 'unfold' and 'rewire'.

Our analysis has automatically reidentified many of the known regulators and their correct function, has suggested additional roles for at least 175 more regulators not associated before with hematopoiesis and has identified points in the lineage at which regulators are recruited to control a specific gene program or lose their regulatory function. Our ability to automatically reidentify many known regulators at the appropriate developmental stage and the significant correspondence among the predicted regulators, known functions, enrichment for *cis*-motifs and enrichment by ChIP followed by deep sequencing supports the probably high quality of our new predictions. Among those, we experimentally tested and confirmed a previously unknown role for ETV5 in the differentiation of the γδ T effector cell subset. Additional studies should determine whether ETV5 regulates the differentiation of IL-17-producing γδ effector cells by selectively controlling the expression of genes in γδ lineage–specific modules.

Ontogenet's rich model allows us 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. The ability to identify regulators that act only during specific differentiation windows helps to detect 'early' programming transcription factors whose expression is shut off when cells transit to the mature stage. However, integrating across the model's predictions in an entire lineage helps to identify transcription factors 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 differentiation steps at which regulatory control 'rewires' most substantially and the regulators that control such 'rewiring'.

As with all expression-based methods used 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 autoregulation in other mammalian circuits that controls cell states[3,25] suggest that such regulatory interactions are probably functional, although some may be 'false-positive' results. Conversely, the activation of other functional regulators may not be reflected by their expression, and some may have been filtered by our stringent criteria (for example, *Tal1*, which encodes a known HSC regulator). Those 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 unique expression profiles that are assigned to modules with similar but distinct expression profiles (such as *Rag1* and *Rag2* in module C5). The inferred regulatory program is unlikely to hold true for those genes.

A similar study of human hematopoiesis[3] has suggested substantial mixed use of modules by lineages, whereas the mouse compendium suggests that most modules are lineage specific. As has been shown before, global profiles, lineage-specific signatures and gene-coexpression patterns are otherwise broadly conserved between humans and mice[4]. One possible reason for the diminished 'mixed use' in the mouse program is that whereas the mouse data set contains many more cell types, it does not include erythrocytes, megakaryocyte, basophils and eosinophils, the cells for which many of the 'mixed-use' patterns have been observed in humans[3]. Notably, many regulators were shared across lineages. In particular, some regulators were active in only one lineage in some modules but were shared by lineages in other modules. For example, ATF6 was an activator in all lineages in the myeloid modules C25, C45 and C49 but was a

T cell–specific repressor in the T cell precursor module C57 and was a T cell–specific activator in the B cell module C71.

Ontogenet is applicable to other differentiation data sets, including data obtained with fetal samples or for cancer studies, when other predictors are used as candidate regulators (for example, genetic variants as in Lirnet[6]), when cells are measured in both the resting state and stimulated state, or for protein-expression data (for example, single-cell, high-dimensional phosphoproteomic mass cytometry data[26]). In each case, the ability to share regulatory programs by related cell types or conditions can both enhance the power and help with biological interpretation. Notably, Ontogenet now depends on a preconstructed ontogeny. Although much is known about the hematopoietic lineage, some parts remain unstructured (for example, all DCs in the myeloid lineage) and some progenitors are not known (for example, those of 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 reflects in part simply the present lack of knowledge of the particular progenitor of a given cell type. New methods would be needed to construct an ontogeny automatically or to revise an existing one. In other cases, Ontogenet's output can be used to refine the topology of the ontogeny 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 identified regulators and regulatory relations are all available for interactive searching and browsing and for downloading at the ImmGen portal and will provide an invaluable resource for future studies of the role of gene regulation in cell differentiation and immunological disease.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession code. GEO: microarray data, GSE15907.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the ImmGen core team (including M. Painter and S. Davis) for help with data generation and processing; eBioscience, Affymetrix and Expression Analysis for support of ImmGen; L. Gaffney for help with figure preparation and layout of the lineage tree; S. Hart for initial layout of the lineage tree; and A. Liberzon (Molecular Signatures Database) for the positional gene sets for mouse. Supported by National Institute of Allergy and Infectious Diseases (R24 AI072073 to the ImmGen Consortium), the US National Institutes of Health (A.R.; and U54-CA149145 and 149644.0103 to V.J. and D.K.), the Burroughs Wellcome Fund (A.R.), the Klarman Cell Observatory (A.R.), the Howard Hughes Medical Institute (A.R.), the Merkin Foundation for Stem Cell Research at the Broad Institute (A.R.) and the National Science Foundation (DBI-0345474 to V.J. and D.K.).

AUTHOR CONTRIBUTIONS

V.J., T.S., A.R. and D.K. designed the study; V.J. developed the algorithm, with input from T.S., A.R. and D.K.; T.S. analyzed the data; K.S. did experiments; O.Z. provided the motif-related code and participated in writing; X.S. generated a mouse model; J.K. designed the experiments and participated in writing the manuscript; and V.J., T.S., A.R. and D.K. wrote the manuscript with input from all authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Heng, T.S.P. *et al.* The Immunological Genome Project: networks of gene expression in immune cells. *Nat. Immunol.* **9**, 1091–1094 (2008).
2. Iwasaki, H. & Akashi, K. Myeloid lineage commitment from the hematopoietic stem cell. *Immunity* **26**, 726–740 (2007).
3. Novershtern, N. *et al.* Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell* **144**, 296–309 (2011).
4. Shay, T. *et al.* Conservation and divergence in the transcriptional programs of the human and mouse immune systems. *Proc. Natl. Acad. Sci. USA* **110**, 2946–2951 (2013).
5. Kim, H.D., Shay, T., O'Shea, E.K. & Regev, A. Transcriptional regulatory circuits: predicting numbers from alphabets. *Science* **325**, 429–432 (2009).
6. Lee, S.-I. *et al.* Learning a prior on regulatory potential from eQTL data. *PLoS Genet.* **5**, e1000358 (2009).
7. Segal, E. *et al.* Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34**, 166–176 (2003).
8. Capaldi, A.P. *et al.* Structure and function of a transcriptional network activated by the MAPK Hog1. *Nat. Genet.* **40**, 1300–1306 (2008).
9. Ram, O. *et al.* Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. *Cell* **147**, 1628–1639 (2011).
10. Miller, J.C. *et al.* Deciphering the transcriptional network of the dendritic cell lineage. *Nat. Immunol.* **13**, 888–899 (2012).
11. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.* **67**, 301–320 (2005).
12. Tamura, T. *et al.* IFN regulatory factor-4 and -8 govern dendritic cell subset development and their functional diversity. *J. Immunol.* **174**, 2573–2581 (2005).
13. Orkin, S.H. & Zon, L.I. SnapShot: hematopoiesis. *Cell* **132**, 712.e711–712.e712 (2008).
14. Pierre, M., Yoshimoto, M., Huang, L., Richardson, M. & Yoder, M.C. VEGF and IHH rescue definitive hematopoiesis in Gata-4 and Gata-6–deficient murine embryoid bodies. *Exp. Hematol.* **37**, 1038–1053 (2009).
15. Kishi, H. *et al.* Lineage-specific regulation of the murine RAG-2 promoter: GATA-3 in T cells and Pax-5 in B cells. *Blood* **95**, 3845–3852 (2000).
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. *Eur. J. Immunol.* **37**, 884–895 (2007).
17. Ji, H. *et al.* Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature* **467**, 338–342 (2010).
18. Lee, J.-W. *et al.* DACH1 regulates cell cycle progression of myeloid cells through the control of cyclin D, Cdk 4/6 and p21Cip1. *Biochem. Biophys. Res. Commun.* **420**, 91–95 (2012).
19. Ng, S.S.M., 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 Lett.* **585**, 1331–1337 (2011).
20. Wang, T. *et al.* Inhibition of activation-induced death of dendritic cells and enhancement of vaccine efficacy via blockade of MINOR. *Blood* **113**, 2906–2913 (2009).
21. Neumann, M. *et al.* Differential expression of Rel/NF-κB and octamer factors is a hallmark of the generation and maturation of dendritic cells. *Blood* **95**, 277–285 (2000).
22. Outram, S.V. *et al.* KLF13 influences multiple stages of both B and T cell development. *Cell Cycle* **7**, 2047–2055 (2008).
23. Yamada, T., Park, C.S., Mamonkin, M. & Lacorazza, H.D. Transcription factor ELF4 controls the proliferation and homing of CD8+ T cells via the Kruppel-like factors KLF4 and KLF2. *Nat. Immunol.* **10**, 618–626 (2009).
24. Narayan, K. *et al.* Intrathymic programming of effector fates in three molecularly distinct γδ T cell subtypes. *Nat. Immunol.* **13**, 511–518 (2012).
25. Yosef, N. & Regev, A. Impulse control: temporal dynamics in gene transcription. *Cell* **144**, 886–896 (2011).
26. Bendall, S.C. *et al.* Single-cell Mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332**, 687–696 (2011).

The complete list of authors is as follows: Adam J Best[9], Jamie Knell[9], Ananda Goldrath[9], Vladimir Jojic[1], Daphne Koller[1], Tal Shay[2], Aviv Regev[2,5], Nadia Cohen[11], Patrick Brennan[11], Michael Brenner[11], Francis Kim[12], Tata Nageswara Rao[12], Amy Wagers[12], Tracy Heng[13], Jeffrey Ericson[13], Katherine Rothamel[13], Adriana Ortiz-Lopez[13], Diane Mathis[13], Christophe Benoist[13], Natalie A Bezman[14], Joseph C Sun[14], Gundula Min-Oo[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], Paul Monach[17], David A Blair[18],

Michael L Dustin[18], Susan A Shinton[19], Richard R Hardy[19], David Laidlaw[20], Jim Collins[21], Roi Gazit[22], Derrick J Rossi[22], Nidhi Malhotra[3], Katelyn Sylvia[3], Joonsoo Kang[3], Taras Kreslavsky[23], Anne Fletcher[23], Kutlu Elpek[23], Angelique Bellemare-Pelletier[23], Deepali Malhotra[23] & Shannon Turley[23]

[9]Division of Biological Sciences, University of California San Diego, La Jolla, California, USA. [10]Broad Institute and Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [11]Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Boston, Massachusetts, USA. [12]Joslin Diabetes Center, Boston, Massachusetts, USA. [13]Division of Immunology, Department of Microbiology & Immunobiology, Harvard Medical School, 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]Department of Medicine, Boston University, Boston, Massachusetts, USA. [18]Skirball Institute of Biomolecular Medicine, New York University School of Medicine, New York, New York, USA. [19]Fox Chase Cancer Center, Philadelphia, Pennsylvania, USA. [20]Computer Science Department, Brown University, Providence, Rhode Island, USA. [21]Department of Biomedical Engineering, Howard Hughes Medical Institute, Boston University, Boston, Massachusetts, USA. [22]Program in Molecular Medicine, Children's Hospital, Boston, Massachusetts, USA. [23]Dana-Farber Cancer Institute and Harvard Medical School, Boston, Massachusetts, USA.

ONLINE METHODS

Data set. Expression of mouse genes was measured on Affymetrix Mogen1 arrays (Affymetrix annotation version 31). Sorting strategies for the ImmGen populations are available on the ImmGen website (http://www.immgen.org/Protocols/ImmGen Cell prep and sorting SOP.pdf).

As the data set of the ImmGen gradually grew from 2010 to 2012, clustering, regulatory program reconstruction and final presentation were done on three different ImmGen releases (September 2010, March 2011 and April 2012) with attempts 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 done on the ImmGen release of September 2010, with 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 membership in clusters from the earlier analysis but used only some of the samples to learn the regulatory program. The heat maps presented here include 755 samples (244 cell types), excluding control samples. For simplicity, only 720 samples are 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 heat maps and tree. The ImmGen website is continuously updated.

Data preprocessing. Expression data were normalized as part of the ImmGen pipeline by the robust multiarray average method. Data were $\log_2$ transformed. For genes with more than one probe set on the array, only the probe set with the highest mean expression was retained. Of those, only probe sets with a s.d. value above 0.5 for the entire data set were used for the clustering, which resulted with 7,965 unique genes with a difference in expression in the September 2011 release and 8,431 in the April 2012 release.

Lineage-specific signatures. We calculated signatures for 11 lineages: granulocyte, macrophage, monocyte, DC, B cell, NK cell, $CD4^+$ T cell, $CD8^+$ T cell, NKT cell, γδ T cell and stem and progenitor cell. Assignment of samples into lineages is in **Supplementary Table 2**. One-way analysis of variance was done for each of the 6,997 genes with an expression value above $\log_2(120)$ in at least one lineage, followed by post-hoc analysis (functions anova1 and multcompare in MATLAB software). For each of the 11 lineages, a gene was considered induced if it had significantly higher expression in that lineage than in all other lineages. A gene was considered repressed if it had significantly lower expression in that lineage than in all other lineages. A false-discovery rate (FDR) of 10% was applied to the analysis of variance *P* values of all genes.

Definition of modules. Modules were defined by clustering. For coarse-grained modules, clustering was done by superparamagnetic clustering (SPC)[27], a principled approach for choosing stable clusters from a hierarchical setting. SPC was used because it does not require a predefined number of clusters but instead identifies the number inherently supported by the data. The clusters defined by SPC are stable across a range of parameters, although they can have variable degrees of compactness. SPC was run with default parameters, which resulted in 80 stable clusters (coarse-grained modules C1–C80); the remaining unclustered genes were grouped into a separate cluster (C81).

Each coarse-grained module was further partitioned into fine-grained modules by affinity propagation clustering[28], with correlation as the affinity measure. The 'self-responsibility' parameter (which 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 clustering of all genes, because it must work with a 'sparsified' affinity matrix.

Clustering resulted in 334 fine-grained modules (F1–F334). On average, 3.9 fine-grained modules were nested in a single coarse-grained module. The minimum number of fine modules nested in a coarse-grained module was 1 (23 coarse-grained modules) and the maximum was 11 (7 coarse-grained modules).

Choice of candidate regulators. Candidate regulators were curated from the following sources: mouse orthologs of all the genes encoding molecules used as candidate regulators in a published study of human hematopoiesis[3]; genes annotated with the gene-ontology term 'transcription factor activity' in mouse, human or rat; genes for which there is a known DNA-binding motif in TRANSFAC matrix database (version 8.3)[29], the JASPAR database (version 2008)[30] and experimentally determined position weight matrices (PWMs)[31,32]; and genes with published data obtained by ChIP followed by deep sequencing or ChIP followed by microarray (**Supplementary Table 11**). Regulators that were not measured on the array or whose expression did not change sufficiently (s.d. < 0.5 across the entire data set) 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 members of the ImmGen Consortium. Each group created its own sublineage tree, and the sublineage trees were connected on the basis of the best knowledge available at present, although some edges are hypothetical (dashed lines, **Fig. 1**). There are two roots to the tree: long-term stem cells from adult bone marrow, and long-term stem cells from fetal liver. 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 were assumed to have a common progenitor or common regulatory program but for which this hypothetical population was not measured (for example, granulocytes 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 (**Supplementary Table 13**).

Module regulatory program. Ontogenet takes the following as input: gene-expression profiles across many different cell types; a partitioning of the genes into modules (the coarse-grained and fine-grained clusters described above); a predefined set of candidate regulators; and 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's 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 i, a member of module m, in cell type t as

$$X_{i,t} = \sum_r w_{m,r,t} a_{r,t} + \varepsilon_{m,t}$$

where each $\varepsilon_{m,t}$ is a Gaussian random variable with 0 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 combination of module, regulator and cell type. Because of parameter tying enforced by the model, the effective number of parameters is much 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

doi:10.1038/ni.2587

member genes across all replicates of the cell type. Although 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 ten 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 because of 'over-parameterization' of the problem, it must be 'regularized' with an extension of the fused Lasso framework[33], which gives rise to a penalized regression problem of the form

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

where $J(w)$ is a chosen penalty. In our case, this penalty is composed of two parts, one that promotes sparsity and selection of correlated predictors and another that promotes 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 $\Sigma_m \Sigma_r \Sigma_t \mid w_{m,r,t} \mid$. However, this penalty tends to be overly aggressive in inducing sparsity and thus prunes many highly correlated predictors and selects only a single representative. Such aggressive pruning may be inappropriate, as the correlated regulators may all be biologically relevant because of 'redundancy' in densely interconnected regulatory circuits. That can be counteracted by the addition of squared terms $\frac{1}{2}\Sigma_m\Sigma_r\Sigma_t(w_{m,r,t})^2$, which yields a composite penalty known as 'Elastic Net'[11], as proposed before[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

$$\lambda\,||w_m||_1+\tfrac{\kappa}{2}||w_m||_2^2$$

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, $\Sigma_r\,|w_{m,r,t_1}-w_{m,r,t_2}|$. The key observation is that $|w_{m,r,t_1}-w_{m,r,t_2}|=0$ 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 $\Sigma_{(t_1,t_2)\in f}\Sigma_r\,|w_{m,r,t_1}-w_{m,r,t_2}|$. 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, with 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

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

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 and 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 λ, κ and δ was geometric, e^{-7}, e^{-6},…, e^3 spanning values between 0.001 and 20) and chose the optimal set of parameters with the Bayesian information criteria (described below).

To simplify the discussion of the optimization, we introduce the sparse predictor matrix A of size (RT) × (T), where $A_{t,(r-1)T+t}=a_{r,t}/\sigma_{mt}$ and = 0 otherwise. Furthermore, we note that the optimal w_m depends only on the mean expression profile of the module's genes and we can introduce variable $y_t=\frac{1}{\sigma_{mt}}\Sigma_{i\in m}\frac{1}{n_m}x_{i,t}$. Hence we can rewrite the objective as

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

Finally we can absorb the term $\frac{\kappa}{2}||w_m||_2^2$ 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 program similar to that of the coarse-grained module in which they are nested. This is accomplished by the 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

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

where the last term ties the programs of the coarse-grained and fine-grained modules. 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 regulatory-program-fitting problems for both coarse-grained and fine-grained module 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 that optimization problem by adding variables that decouple the penalties:

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

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

Model selection with Bayesian information criterion. The formulation of our optimization problem above is dependent on the set of parameters λ, κ and δ; we obtain a model by solving the convex problem above for a particular combination of λ, κ and δ. Different combinations of these parameters will yield regulatory programs of different quality. One way to identify the optimal λ, κ and δ is through the use of held-out data or through cross-validation. However, a search for these parameters with cross-validation would be 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.

This 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

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

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. To make this more formal, we consider matrix A and construct its counterpart, B. We use $A_{r,t}$ to denote a column of matrix A. We then construct a graph in which nodes correspond to columns of matrix A. Given two nodes corresponding to A_{r,t_1} and A_{r,t_2}, the graph will have an edge between these two nodes if cell type t_1 is a parent of cell type t_2, and $w_{m,r,t_1} = w_{m,r,t_2}$. 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 the number of columns of A in the connected component associated with a column of B.

Hence we can compute the $\mathrm{BIC}(w)$ as

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

Post-processing of regulatory programs. Once we obtain an optimal regulatory program in terms of BIC, we use post-processing to remove regulatory relationships for underexpressed regulators. We placed a low expression cutoff of 5.5 on the $\log_2$ scale. At this cutoff, 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 regulators. For each lineage-specific module, we automatically queried its regulators in PubMed with the name of the lineage(s). For each module that was upregulated or downregulated with differentiation, we queried its regulators with 'hematopoietic differentiation'. All PubMed queries were done on 30 July 2012.

Choice of lineage regulators. For each lineage, we collected the regulators deemed active by having a nonzero activity weight and significant expression in excess of 9.0 on the $\log_2$ 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 on the basis of 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 expression were ranked higher than were regulators that were infrequently active or had low expression. The regulators with the highest expression typically were given the highest total activity weight across lineages.

Notably, this procedure, although straightforward, will not reflect all the lineage regulators identified by the model. First, those lineage regulators that act only during a limited window (for example, early in differentiation) would be under-represented by this analysis yet would be captured in the overall model in the window in which they act. Second, because of the post-processing step (described above), regulators with high baseline expression can have a constant activity weight even if their expression is very lineage specific (for example, GATA-3) and thus be under-represented in the recruitment analysis (although 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 genome browser website of the University of California Santa Cruz (http://hgdownload.cse.ucsc.edu/downloads.html). For each gene, we scanned the region starting from position –1,000 (base pairs upstream of the transcription start site) and ending at position +200 (base-pairs downstream of the transcription start site). We represented the nucleotide at position j (relative to –1,000 bp from the transcription start site) for gene i as $S_{i,j}$. We represented each *cis*-regulatory element by a PWM. We compiled a set of 1,651 PWMs from the TRANSFAC matrix database (version 8.3)[29], the JASPAR database (version 2008)[30] and experimentally determined PWMs[31,32]. We denote the PWM of the 'k-th' motif 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 (that is, 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 $\mathrm{LOD}(i,j,k)$, defined as the log likelihood ratio (LOD score) for observing the sequence given the PWM versus a given random genomic background:

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

Genomic background was determined as $P_b(\text{'A'}) = P_b(\text{'T'}) = 0.3$, $P_b(\text{'C'}) = P_b(\text{'G'}) = 0.2$, which represents the nucleotide composition of the mouse genome. We then found the best motif instance over the entire promoter region, defined as $MAX\text{-}LOD(i,k) = max_j LOD(i,j,k)$.

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

$$IC(k) = L_k + \sum_{i=1}^{L_k}\sum_{j=1}^{4} P_k(i,j)\log_2 P_k(i,j)$$

We defined the PWM-specific threshold for the 'k-th' motif as τ_k, the $1 - 2^{-\mathrm{IC}(k)}$ quantile of the PWM LOD distribution across all genes' promoters. We considered a 'hit' for the 'k-th' motif at the 'i-th' gene if the best score ($MAX\text{-}LOD(i,k)$) 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 relative to that of the entire set of genes assigned to modules serving as background. An enrichment of a motif in a module results in higher than expected $MAX\text{-}LOD$ scores for the genes in this module; to capture this effect, we computed the P value by comparing the scores $MAX\text{-}LOD(i,k)$ for all genes i in the module M and the scores for the entire set of genes assigned to modules by a one-sided rank-sum test. We then used an FDR of 5% on the entire matrix of P values $p_e(M,k)$ and declared all P values that passed this procedure significant 'hits'. The FDR was calculated separately for coarse-grained and fine-grained modules.

Binding events enrichment. The public data sets obtained by ChIP followed by deep sequencing and ChIP followed by microarray (**Supplementary Table 11**) were downloaded from the GEO (Gene Expression Omnibus) database repository, supplementary material and designated sites in the original publications (**Supplementary Table 11**; 360 experiments of 109 unique regulators). The target list defined in each original publication was used whenever available. Otherwise, genes that had a binding event reported from the position 1,000 base pairs upstream of the transcription start site to the position 200 base pairs downstream of the transcription start site were listed as targets. In data sets obtained with human samples, gene symbols were replaced by the mouse gene symbol wherever a one-to-one ortholog exists according to the phylogenetic resource EnsemblCompara[35]. Only genes included in the clustering were considered targets for the purpose of the calculation of enrichment.

The hypergeometric P value was calculated for the size of intersection of each module with each target list. An 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.

doi:10.1038/ni.2587

Estimating the 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 for which the 'universe sizes' were the number of possible regulatory interactions including the overlapping regulators. For example, for estimation of the significance of the overlap of ChIP and Ontogenet regulatory interactions, the 'universe size' is the number of regulators that were candidates for Ontogenet and had ChIP information multiplied by the number of modules. The ChIP interactions are the enriched modules according to the ChIP data set, 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 last *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 differ in significance for the three-method 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. Curated gene sets (C2), motif gene sets (C3) and gene ontology (GO) gene sets (C5) from the Molecular Signatures Database (version V.3) were downloaded from the Broad Institute website (http://www.broadinstitute.org/gsea)[36]. For each group, gene symbols were replaced by the mouse gene symbol wherever a one-to-one ortholog exists according to EnsemblCompara. Only genes included in the clustering were considered functional group members for the purpose of the calculation of enrichment.

A hypergeometric *P* value was calculated for the size of intersection of each module with each functional group. An 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 and C5).

Identification of differentiation steps with a change in activity weight of regulators. For each module and each edge (differentiation step) of the hematopoietic tree, the activity weight of the 'parent' was compared with the activity weight of the 'child', which resulted in one of the following classifications: no change (activity weights are the same); activator recruitment (parent activity weight is 0; child activity weight is positive); activator strengthening (parent activity weight is positive and is smaller than that of the child); activator disposal (parent activity weight is positive and child activity weight is 0); repressor recruitment (parent activity weight is 0; child activity weight is negative); repressor strengthening (parent activity weight is negative and is larger than that of the child); repressor disposal (parent activity weight is negative and child activity weight is 0). For simplicity, we omitted the 'regulator weakening' option. Those lineage-specific regulators that are assigned constant activity weight across all cell types (such as GATA-3) will not be captured by this analysis but are part of the model.

Mice. Mice with *lox*P-flanked *Etv5* alleles (*Etv5*$^{fl/fl}$)[37] were crossed with C57BL/6 mice with a transgene encoding Cre recominase driven by the promoter of the gene encoding CD2 to generate mice with T cell–specific ETV5 deficiency (CD2p-CreTg^{+}*Etv5*$^{fl/fl}$, backcrossed three times to the C57BL/6 strain). The *lox*P-flanked *Etv5* locus is specifically deleted from the genome starting in CD25^{+}CD44^{-}CD3^{-}CD4^{-}CD8^{-} thymic precursors (DN3) with ~80% deletion efficiency in γδ thymocyte subsets, as inferred from the analysis of Cre-activity-reporter mice (CD2p-CreTg^{+}Rosa-STOP$^{fl/fl}$-EYFP).

Flow cytometry. Intracellular staining (Cytofix/Cytoperm Kit; BD Biosciences) and intranuclear staining (FoxP3 Staining Kit; eBioscience) were done as described[24]. The following antibodies were used: anti-TCRδ (GL3), anti-CD24 (HSA, M1/69), anti-CD44 (IM7), anti-CD62l (MEL-15), anti-IL-17A (ebio17B7) and anti-RORγt (AFKJS-9; all from eBioscience); and anti-$V_{\gamma}2$ (UC3-10A6), anti-$V_{\delta}6.3$ (8F4H7B7), anti-CCR6 (140706) and anti-CD27 (LG.3A10; all from BD Biosciences). Anti-$V_{\gamma}1.1$ (2.11) was purified from culture supernatant and was biotinylated with the FluoReporter Mini-Biotin-XX Labeling Kit (Invitrogen). Data were acquired on an LSRII (BD) and were analyzed with FlowJo software (Treestar).

27. Blatt, M., Wiseman, S. & Domany, E. Superparamagnetic clustering of data. *Phys. Rev. Lett.* **76**, 3251–3254 (1996).
28. Frey, B.J. & Dueck, D. Clustering by passing messages between data points. *Science* **315**, 972–976 (2007).
29. Matys, V. *et al.* TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108–D110 (2006).
30. Sandelin, A., Alkema, W., Engström, P., Wasserman, W.W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32** (suppl. 1), D91–D94 (2004).
31. Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720–1723 (2009).
32. Berger, M.F. *et al.* Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **133**, 1266–1276 (2008).
33. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* **67**, 91–108 (2005).
34. Boyd, S.P. & Vandenberghe, L. in *Convex Optimization*, Ch 11.7 (Cambridge University, Cambridge, UK, 2004).
35. Vilella, A.J. *et al.* EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* **19**, 327–335 (2009).
36. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
37. Zhang, Z., Verheyden, J.M., Hassell, J.A. & Sun, X. FGF-regulated Etv genes are essential for repressing Shh expression in mouse limb buds. *Dev. Cell* **16**, 607–613 (2009).