Conservation and divergence in modules of the transcriptional programs of the human and mouse immune systems*

Tal Shay[1,2]‡, Vladimir Jojic[3,4], Or Zuk[1,5], Christophe Benoist[6], Daphne Koller[3], Aviv Regev[1,7]‡ and ImmGen consortium[#]

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

[2]Department of Life Sciences, Ben-Gurion University of the Negev, Beer-Sheva 8410501, Israel (current address)

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

[4]Calico Labs, 1170 Veterans Blvd, South San Francisco, CA 94080 (current address)

[5]Department of Statistics, The Hebrew University of Jerusalem, Jerusalem 91905, Israel (current address)

[6]Department of Microbiology and Immunology, Harvard Medical School, Boston, MA 02115

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

‡ To whom correspondence should be addressed: aregev@broadinstitute.org (AR) talshay@bgu.ac.il (TS).

* This paper describes a complementary analysis to our PNAS paper, Shay, Jojic, et al. (2013). "Conservation and divergence in the transcriptional programs of the human and mouse immune systems." PNAS 110(8): 2946-2951.)

[#]**ImmGen Consortium Members** (as of 2013)

Paul Monach[a], Susan A. Shinton[b], Richard R. Hardy[b], Radu Jianu[c], David Laidlaw[c], Jim Collins[d], Roi Gazit[e], Brian S. Garrison[e], Derrick J. Rossi[e], Kavitha Narayan[f], Katelyn Sylvia[f], Joonsoo Kang[f], Anne Fletcher[g], Kutlu Elpek[g], Angelique Bel- lemare-Pelletier[g], Deepali Malhotra[g], Shannon Turley[g], Adam J. Best[h], Jamie Knell[h], Ananda Goldrath[h], Vladimir Jojic[i], Daphne Koller[i], Tal Shay[j], Aviv Regev[j], Nadia Cohen[k], Patrick Brennan[k], Michael Brenner[k], Taras Kreslavsky[k], Natalie A. Bezman[l], Jo- seph C. Sun[l], Charlie C. Kim[l], Lewis L. Lanier[l], Jennifer Miller[m], Brian Brown[m], Miriam Merad[m], Emmanuel L. Gautier[m,n], Claudia Jakubzick[m], Gwendalyn J. Randolph[m,n], Francis Kim[o], Tata Nageswara Rao[o], Amy Wagers[o], Tracy Heng[p], Michio Painter[p], Jeffrey Ericson[p], Scott Davis[p], Ayla Ergun[p], Michael Mingueneau[p], Diane Mathis[p], and Christophe Benoist[p]

[a]Department of Medicine, Boston University, Boston, MA 02118; [b]Fox Chase Cancer Center, Philadelphia, PA 19111; [c]Computer Science Department, Brown University, Providence, RI 02912; [d]Department of Biomedical Engineering, Howard Hughes Medical Institute, Boston University, Boston, MA 02215; [e]Immune Diseases Institute, Children's Hospital, Boston, MA 02115; [f]Department of Pathology, University of Massachusetts Medical School, Worcester, MA 01655; [g]Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA 02215; [h]Division of Biological Sciences, University of California at San Diego, La Jolla, CA 92093; [i]Computer Science Department, Stanford University, Stanford, CA 94025; [j]Broad Institute, Cambridge, MA 02141; [k]Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Boston, MA 02467; [l]Department of Microbiology and Immunology, University of California, San Francisco, CA 94143; [m]Icahn Medical Institute, Mount Sinai Hospital, New York, NY 10029; [n]Department of Pathology and Immunology, Washington University, St. Louis, MO 63110; [o]Joslin Diabetes Center, Boston, MA 02215; and [p]Division of Immunology, Department of Microbiology and Im- munobiology, Harvard Medical School, Boston, MA 02115

Abstract

Studies in mouse have shed important light on human hematopoietic differentiation and disease. However, substantial differences between the two species often limit the translation of findings from mouse to human.

Here, we compare previously defined modules of co-expressed genes in human and mouse immune cells based on compendia of genome-wide profiles. We show that the overall modular organization of the transcriptional program is conserved. We highlight modules of co-expressed genes in one species that dissolve or split in the other species. Many of the associated regulatory mechanisms – as reflected by computationally inferred *trans* regulators, or enriched *cis*-regulatory elements – are conserved between the species. Nevertheless, the degree of conservation in regulatory mechanism is lower than that of expression, suggesting that distinct regulation may underlie some of the conserved transcriptional responses.

Introduction

The process of differentiation from hematopoietic stem cells into all blood and immune cells is extensively studied in mammals, but the transcriptional circuitry controlling this process is still only partially understood. Studies of human immunology typically focus on cells isolated from peripheral blood, bone marrow or cord blood, with *in vivo* studies mostly monitoring clinical manipulations, such as vaccinations or transplantations. Conversely, studies on mice, including in immune-deficient mice transplanted with human hematopoietic cells, can be done in-vivo[1]. Moreover, all major immune cell populations are shared by human and mouse. Therefore, the mouse is an important model organism for human immunology.

However, there are substantial differences between the two species, and studies have repeatedly suggested caution when translating findings from mouse to human[2-5]. Biologically, these differences include: life span, total number of cell divisions[6], several different immune cell markers[7], distinct physiological phenotypes caused by deficiencies in the 'same' (orthologous) gene in the two species (*e.g.*, MYD88[8], STAT5B[9]), and genes that are present in only one of the two species, usually due to species- or clade-specific expansion or contraction of multi-gene families[7]. There are also significant differences arising from experimental protocols: mouse studies are primarily conducted in inbred mouse strains that are controlled for age, environmental exposures and other confounding factors, in contrast to natural human populations; some mouse models of human disease and therapy do not mimic well human clinical applications (*e.g.*, asthma[10]); in some cases, the markers used to define the 'same' cell populations are shifted in their expression during differentiation in the two species (*e.g.*, in hematopoietic stem cells[3]);

and cells analyzed in mice span the range of lymphoid organs, while studies of human material are usually limited to blood.

Studies in a variety of species and cell types have shown that transcriptional profiles provide a detailed view into the molecular and functional state of cells, help identify relevant regulatory mechanisms[11, 12], and pinpoint important variations between species[13-16]. In particular, comparing modular composition and regulatory programs may provide important insights into both conservation and divergence of the immune system. Whereas the classical orthologous transcription factors are known to play conserved key roles in both human and mouse immune system differentiation (*e.g.*, RUNX1 and TAL1 in hematopoietic stem cells, PAX5 and EBF1 in B-cells, TCF1 and GATA3 in T-cells), recent studies in other cell types suggest that as many as 50% of the interactions between proteins coding for transcription factors may not be conserved between human and mouse[17], and that there is a fast turnover of actual binding of transcription factors to DNA between human and mouse[18] and between other vertebrate species[19]. In particular, it has been estimated that 32%–40% of the functional regulatory sites in human are not functional in rodents[20].

The generation of two unprecedented compendia of gene expression measured in a range of multipotent and differentiated cell types across the human[21] and mouse[22] immune lineages opened the way to systematically decipher the regulatory circuitry controlling hematopoiesis. We previously used these two datasets to show that transcriptional profiles of orthologous lineages are globally similar between the species, signatures of lineage-specific gene expression are shared, and the expression patterns of most genes are conserved between the species[23]. We

highlighted genes with different expression patterns in human and mouse that were not previously reported and validated a few of them experimentally[23]. Mapping modules of co-expressed genes between human and mouse will teach us about the similarities and differences of immune differentiation, and point to processes that are different, or differently regulated.

Here, we compare the modules defined independently in the human and mouse compendia and show that there is a significant similarity in the modular organization of the transcriptional program in both species, and a corresponding significant overlap in the underlying regulatory programs as defined by the inferred active regulators and associated *cis*-regulatory elements in the promoters of target genes.

Results

Transcriptional maps of the human and mouse immune systems

We compared two compendia of the human and mouse immune systems. The human compendium, termed differentiation map (D-map) [21] consists of 211 samples from 38 cell types (4-8 samples per cell type), measured on Affymetrix U133A arrays (22,268 probesets). The mouse Immunological Genome (ImmGen) compendium [22] consists of 802 samples from 244 immune cell types (~3 samples per cell type), measured on Affymetrix MoGene 1.0 ST arrays (25,194 probesets excluding control and unassigned features). We mapped 10,248 one-to-one orthologs between the two species that were measured in both studies (**Methods**).

There are several distinctions in the design of the two studies. However, gene expression across the common lineages [Hematopoietic stem and progenitor cell (HSPC), granulocytes (GN), monocytes (MO), dendritic cells (DCs), B-cells, natural killer (NK) cells, and T-cells] was shown to be similar[23].

Significant conservation of co-expression in most modules, especially those highly expressed in stem and progenitor cells

We compared the overall organization of the transcriptional programs in the two species. Studies from yeast[24] to human[11] – including of the human hematopoiesis compendium[21] – have shown that transcriptional programs are organized in a modular fashion. In some cases, modules of co-expressed genes are conserved across species, even kingdoms[25, 26]. In principle, modules can be conserved at several levels, including (**1**) conservation of gene membership

(orthologous genes assigned to the 'same' co-expression module); (**2**) conservation of both membership and expression pattern (orthologous genes with conserved expression profiles under comparable conditions); or (**3**) conservation of a part ('core') of a module in both species (membership and/or expression), with additional genes in the module in one species, but not the other (and vice versa).

Focusing first on conservation of co-expression ((**1**) above), we examined modules of co-expressed genes defined independently for each compendium in its entirety (**Methods**, [21, 27]). Since noise in modules reconstructed independently may somewhat reduce (but not increase) the degree of observed conservation, we assessed – for each module defined in one species (*e.g.*, human) – the degree of co-expression of its members' orthologs in the other species (*e.g.*, mouse), independent of their module assignment in that other species (*e.g.*, mouse) (**Fig. 1**). To this end, we used a $Z_{summary}$ score (**Methods**) of module preservation, previously suggested for use across datasets and species[28]. The $Z_{summary}$ score combines several measures of module conservation to reflect, for each module, the degree of co-expression of its members' orthologs in the other species, and their distinction from genes not in the module. A $Z_{summary} > 2$ indicates significant conservation. The $Z_{summary}$ statistic estimates conservation of co-expression more sensitively than a simple overlap measure between modules (below), which can be biased by the hard partitioning of similarly expressed genes to different modules with somewhat similar expression patterns.

By these measures, we find that 41 of the 80 human modules tested and 23 of the 67 mouse modules tested are conserved ($2 < Z_{summary}$) in mouse or human, respectively (**Fig. 1b and c,**

Supplementary Table 1). In most cases, both the co-expression and the actual expression pattern of the module's genes are conserved, as reflected by the modules conservation of expression (COE) scores, calculated, following[23], as the Pearson correlation coefficient between the mean module expression in the common lineages in both species (mean COE > 0.5 in 33 of 41 human conserved modules and 19 of 23 mouse conserved modules; option (**2**) above). The most highly conserved human module (module 823, $Z_{summary} > 10$) consists of genes whose expression is down-regulated with differentiation, suggesting particular conservation of the stem and progenitor transcriptional program. In eight human modules and four mouse modules only co-expression is conserved, but the actual expression pattern is not (COE < 0.5). Notably, module conservation does not necessarily imply conservation of the underlying regulatory program. Indeed, there is substantial evidence in microorganisms [29, 30] that co-expressed modules can be conserved across species, even when their underlying regulatory mechanisms diverge substantially.

Conservation of the modular organization of the immune transcriptional program

While individual genes may be members of a module in one species but not the other, the overall modular organization of the program is maintained, as reflected by 61 pairs of human and mouse modules with significant overlap (**Fig. 2a and d**, hyper-geometric test, FDR 10%), involving 46 distinct human modules (**Fig. 2b and c**, left) and 31 mouse modules (**Fig. 2b and c**, right), with overlaps ranging from 14% to 80% (**Fig. 2b**, right most bar chart). In most of the significantly overlapping module pairs, the expression pattern of the two modules is highly similar (median COE of module means 0.69; **Fig. 2b,** second right most bar chart, **Supplementary Table 2**). However, there are cases where only the co-expression is conserved, for example the human

module 673, induced in B-cells, and mouse module 32, induced in dendritic cells, which significantly overlap (**Fig. 2c**, bold/asterisks).

In some cases, the lack of one-to-one correspondence between modules is a result of the different cell populations profiled in the two compendia, leading to more refined patterns in one species than in the other. For example, human module #973[21], which consists of genes induced in HSPC and early erythroid progenitors, significantly overlaps with four mouse modules (#7, 13, 14, and 15), which are each induced in HSPC, but differ in expression in specific progenitors and CD8 T-cells, measured only in mouse (**Fig. 2c**, bold/asterisks).

Characterization of divergent modules

Different factors can underlie the 39 (49%) human and 44 (66%) mouse modules that are not conserved by $Z_{summary}$. These include (**Fig. 3a**): (**1**) lack of any co-expression in the other species, including due to differences in cell types between the compendia, resulting in module dissolution; (**2**) a relatively small conserved 'core' accompanied by distinct genes in each species; (**3**) separation into several expression patterns in the other species. In addition, in some cases, the underlying clustering may over- or under-split genes into modules. To test the prevalence of each of these (possibly overlapping) possibilities, we manually inspected the expression pattern of all human modules in mouse, and vice versa.

Four of the human non-conserved modules and 15 of the mouse non-conserved modules are those induced in cell types that are either absent or under-represented in the other species' compendium (option (**1**) above). These include four modules induced specifically in erythroid

progenitors in human (#727, 637, 889, 895), and modules induced in stromal cells (#35-39, 44, 73, 80), T-cell progenitors (#57), specific myeloid cell types (#30-31, 58, 65, 68) or gamma delta T-cells (#56) in mouse. Co-expression of these modules' genes is thus cell-type specific. Additional profiling in the relevant cell types may identify similar co-expression in the corresponding species. Two non-conserved modules cannot be attributed to lack of comparable populations: human modules #859 and 955, which are induced in T-cells, but their orthologous genes in mouse display various expression patterns (**Fig. 3b, Supplementary Fig. 1a**).

In 19 human modules and 21 mouse modules, a 'core' (>50% of the module genes) is conserved (COE > 0.5) between the species (option (**2**) above). In many of these cases, those 'core' genes have higher maximal expression levels than the other genes in the same module (*e.g.*, in mouse modules #18, 25 and 33, **Fig. 3c**, **Supplementary Fig. 1b**), consistent with the higher COE levels of highly expressed genes[23].

In a few cases, the orthologs of a module's members in one species are partitioned into two distinct expression patterns (option (**3**) above), possibly reflecting only a portion of the pattern in the original module. For example, the genes in human modules #649 and 673 are induced in both B-cells and some myeloid cells, whereas their orthologs in mouse are induced in either myeloid cells or in B-cells but in most cases not both (**Fig. 3d, Supplementary Fig. 1c**).

Assessing conservation of regulatory mechanisms in immune cells

Does the evolution of regulatory mechanisms correlate with that of the organization of the transcriptional response in the immune system? In particular, do conserved mechanisms underlie the conserved expression and modules? On the one hand, studies on the evolution of gene

expression in other organisms – especially yeasts – have suggested that even modules with strong conservation of expression[30] and co-expression[29, 31, 32] can be associated with distinct regulatory mechanisms in different species. In mammals, recent studies of transcription factor binding have also demonstrated substantial evolutionary turnover in transcription factor-DNA interactions, even in genes with strong functional and expression conservation[18]. On the other hand, many of the known regulators of human and mouse immune system differentiation have been experimentally demonstrated to conserve their function[33], suggesting that the same regulators orchestrate the process in human and mouse. In particular, turnover in the specific position of a *cis*-regulatory element or physical binding event may exist even when the gene target or biological process controlled by a factor remains conserved.

We previously showed that the expression pattern of most genes encoding transcriptional regulators, including known master regulators, is conserved, a prerequisite to conservation of regulation[23]. To study the extent of conservation of regulatory mechanisms between human and mouse, we examined two aspects of the immune system regulatory program as reflected in the transcriptional compendia and associated modules (**Fig. 4a**): (**1**) inferred *trans*-regulatory associations between regulators and targets as predicted by computational modeling of gene expression patterns; and (**2**) *cis*-regulatory elements enriched in promoters of module members or lineage-signature genes.

Many of the regulators predicted by the model are conserved

In order to identify active *trans* regulators associated with each module, we used Ontogenet, a new method that combines linear regression with the tree structure of the dataset to predict a set of transcriptional regulators that would best account for each module's expression [27]. Ontogenet was already applied to the mouse modules using 580 candidate regulators, and identified 480 regulators (7-45 per module, median 16) [27]. Here we applied Ontogenet to the human modules independently, using 394 candidate regulators in human (**Supplementary Table 3**), and identified 213 regulators in human (0-30 per module, median 15). The larger number of candidate and active regulators in mouse reflects the higher complexity of the mouse dataset. We defined the targets of a regulator as the union of all genes in that regulator's target modules[11]. The regulatory model also associates the regulators with each lineage, in rank order.

Of the 480 regulators chosen by the Ontogenet algorithm in mouse ('active regulators), 155 were also chosen as active regulators in the human model. The COE scores of the 155 active regulators chosen in both models are significantly higher than those of other genes filtered in a similar manner (one sided KS-test, $P=2.7*10^{-6}$), but not of the active regulators that are selected only in one species (one sided KS-test, P=0.12).

Furthermore, the targets and lineages associated with a few of the regulators are significantly conserved between the two regulatory programs. Of the 155 pairs of orthologous regulators chosen in both species, the targets of 13 regulators (CBX5, CIITA, FUBP1, HIVEP2, LRRFIP1, NPM1, PHB2, POU2AF1, SPI1, SPIB, TCF4, TFEC and STAT6) significantly overlap between

the human and mouse models (hyper-geometric test, FDR=10%). Such 13 overlapping pairs are more than would be expected by chance ($P<10^{-3}$, permutation, **Methods**). Finally, many specific regulators are conserved in their association to individual lineages (*e.g.*, **Fig. 4b**, bold/asterisks). For each lineage, we tested the overlap of its activators and repressors between human and mouse (**Supplementary Table 4**). T-cell activators and repressors and DC repressors significantly overlap between the human and mouse models (hyper-geometric test, FDR=10%, **Methods**).

Most of the pairs of modules conserved between human and mouse (40 of 61) are associated with at least one pair of orthologous regulators (**Supplementary Table 2)**. For example, POU2AF1 is a regulator of the B-cell modules human 961 and mouse C33; TBX21 is a regulator of the NK-cell modules human 997 and mouse C19, GATA3 is a regulator of the T cell modules human 667 and mouse C18. Those regulators that were chosen in one species but not the other have a significantly lower range of expression in the species where they were not selected (t-test $P=2.8*10^{-3}$ for mouse; $P=1.2*10^{-4}$ for human**)**. Indeed, many of the discordant regulators are highly expressed in the non-common lineages (*e.g.*, the regulatory T-cell regulator Foxp3 and the stromal specific Epas1, chosen only in mouse, are expressed only in cell types not measured in the human dataset).

Some of the *cis*-regulatory elements associated with modules and signatures are conserved

There is substantial conservation in the *cis*-regulatory elements associated with the regulatory programs in the two species, suggesting conserved regulatory mechanisms. First, 16 of the 61 pairs of modules conserved between human and mouse are associated with at least one *cis*-

regulatory element that is enriched in both members of the pair (**Supplementary Table 2**). For example, human module #973 and its mouse orthologous modules #7, 14, and 15, all consist of genes down-regulated with differentiation, contain many cell-cycle genes, and are enriched with the binding sites of different E2F factors. E2F factors are known cell cycle regulators[34], and E2F1 and E2F3's expression is down-regulated with differentiation.

***Cis*-regulatory conservation is not associated with higher conservation of expression**

Surprisingly, the presence of a conserved *cis*-element is not associated with significantly higher conservation of expression as measured by COE. For example, pairs of orthologous modules that are enriched for the same sequence motif in human and mouse do not have significantly more similar expression than pairs of modules without matched enriched motifs (KS-test P=0.85). This may be either due to the relative paucity of known *cis*-regulatory elements and noise in prediction of their targets[35, 36], or due to the presence of dense regulatory circuits[21], such that loss of *cis*-regulation by one factor is compensated by gain of regulation by another factor active in the same lineage.

Discussion

While many features of the immune system are conserved between human and mouse – making mice a compelling model of human immune system differentiation – there are important known differences between these systems, including in transcriptional profiles (reviewed in Mestas *et al*.[7]). However, previous studies did not systematically analyze the extent of similarity and differences in the human and mouse regulation of transcriptional programs of the immune system. We previously compared two extensive transcriptional compendia, and showed an extensive conservation of the transcriptional program at several levels – global profiles, individual genes and lineage specific gene signatures, and catalogued the transcriptional differences in one to one and one to many orthologs[23].

Here we show conservation and divergence at the level of modules of co-expressed genes, and compare the underlying regulatory mechanisms controlling these programs in the two species. Furthermore, when a regulatory module in one species is partially 'dissolved' in another, co-expression (and hence module membership) of genes with higher expression levels tends to be more conserved, in accordance with previous observation, that the expression pattern of highly expressed genes is more conserved[23].

The most prevalent expression pattern in both species is downregulation with differentiation, and many of the conserved modules contain genes expressed specifically in stem and progenitor cells. This may reflect a stronger purifying selection against changes in the HSPC transcriptional program – where changes could disrupt many differentiation paths – whereas differentiated cell types program may be under less selective pressure.

Module conservation is often reflected by concomitant conservation of the associated regulatory mechanisms. In particular, the expression profiles of most regulators chosen by the regulatory models are well conserved. Indeed, there is a significant overlap between lineage regulators assigned by the two models. Furthermore, many conserved modules are also associated with conserved enrichment of regulatory elements. Nevertheless, in many cases conserved modules are associated with different *cis*-regulatory elements. Such conservation of gene expression ad co-exprtession but divergence of underlying *cis* elements has been previously described in yeast [29, 31, 37]. Some of the discrepancies are likely related to the lower reliability of *cis*-reguatory predictions, or to differences in the sampled cell types. For example, the mouse compendium does not include a counterpart to the interferon gamma-producing CD56+ human NK cells. In other cases, the preponderance of dense regulatory circuits in hematopoiesis[21] may facilitate such divergence here.

Our comprehensive comparison of human and mouse modules of co-expressed genes and their regulation is of great interest to the immunology and evolution community. To facilitate future studies of evolution of genes regulation, we provide all the data and analysis in a separate browser on the ImmGen portal (http://rstats.immgen.org/comparative/comparative_search.php). These should assist researchers in identifying both the broad similarities and fine distinctions in the modular structure of the immune system transcriptional program.

Acknowledgements

We thank the members of the ImmGen Consortium for discussions, the ImmGen core team at the time of analysis (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, Lewis L. Lanier, Roi Gazit, and Paul A. Monach for many helpful discussions, and Noa Novershtern and Benjamin L. Ebert for help with the human dataset.

Figure legends

Figure 1. **Conservation of co-expression in the human and mouse immune system differentiation transcriptional programs.** (**a**) The expression of each gene in a module (*e.g.*, human module 721, left) is compared to that of its ortholog in the other species (*e.g.*, mouse, right). (**b**) Mean expression profiles of all human modules (left heatmap) and of their matching sets of one-to-one orthologs in mouse (right heatmap). Bar chart on the left shows the human module size; grayscale bar in the middle is the $Z_{summary}$ reflecting module conservation (black: $Z_{summary}$ >10 highly conserved; gray: $2<Z_{summary}<10$ conserved; white: $Z_{summary}<2$ not conserved); bar chart on right is the COE between the mean expression of the module and the mean expression of the mouse orthologs of its members. (**c**) Mean expression profiles of all mouse modules (right) and their matching sets of one-to-one orthologs in human (left). Bar chart on left shows the mouse module size; grayscale bar in the middle is the $Z_{summary}$ of the module conservation (gray: $2<Z_{summary}<10$ conserved; white: $Z_{summary}<2$ not conserved); bar chart on right is the COE between the mean expression of the module and the mean expression of the human orthologs of its members. (c) is analogous to (b), but projecting from mouse modules to human genes (c) rather than from human modules to mouse genes (in b).

Figure 2. **Conservation of modularity in the human and mouse immune system differentiation transcriptional programs.** (**a**) The gene membership in each human module (*e.g.*, human module 865, left) and mouse module (*e.g.*, mouse module C40, right) is compared based on orthology (black lines) and the degree of overlap is estimated for significance with a hyper-geometric test (Venn diagram, far right). Shown are the heatmaps of mean expression profiles of the common lineages (**b**) or all cell types (**c**) for the 61 module pairs that significantly overlap between human (left heatmap) and mouse (right heatmap). Module pairs are sorted by human lineage with maximal expression. Bar charts to the left of each expression matrix display module size. Bar charts on the far right are the COE between the mean expression of the matching modules (white bars) and the percentage of genes in the overlap in the human (black) and mouse (gray) modules. The expression patterns in the module pairs are also typically conserved with some exceptions (*e.g.*, human module 673, induced in B-cells, and mouse module C32, induced in dendritic cells). Module pairs discussed in text are marked with bold and asterisks.(**d**) Number of overlapping orthologs (grayscale color bar, right) between every pair of human and mouse modules. Significantly overlapping module pairs are marked by red rectangles.

Figure 3. **Patterns of module conservation and divergence**. (**a**) Schematic visualizations of possible evolutionary patterns when comparing a module in one species (left) to the corresponding one to one orthologs in the other species (right). Top to bottom: conservation, dissolution, partial conservation of a core, and separation into multiple modules. In each case, shown are schematic expression patterns in matching lineages in the two species. (**b**) Human module #955 (left) is dissolved in mouse (right). Genes are sorted by their orthologs' assignment to mouse modules. (**c**) The core of mouse module #18 (right) – induced in T-cells and in some genes also in NKs – is conserved in human (left). Genes are sorted by their mean expression in mouse T-cells. (**d**) Human module #649 (left) – induced in B, GN and MO – is split in mouse (right), where some orthologs are induced in B-cells, and others are induced only in myeloid cells. Genes are sorted by their mean expression in mouse B-cells. In b-d, genes are mean centered (color scale, bottom).

Figure 4. **Regulation of human and mouse immune system differentiation is largely conserved**. (**a**) Schematic of several possible levels of conservation of regulation in human and mouse: expression of regulators (top), regulators association with modules (arrows), and *cis*-regulatory elements enriched in promoters of module genes (rectangles and triangles, right and left). (**b**) Conserved lineage association of regulators. Shown is the lineage tree with selected immune system regulators with conserved expression (COE>0.8). Conserved lineage associations according to the regulatory models are marked with bold and asterisks. The full set of conserved lineage associations is in **Supplementary Table 4**.

Materials and Methods

Datasets and preprocessing

Gene expression in mouse samples was measured on Affymetrix array MoGen, annotation version na31. The ImmGen March 2011 release was used. This release includes 802 arrays with 22,268 probesets, as previously described[27]. Gene expression in D-map human samples was measured on Affymetrix array U133A, annotation version na31, as previously described[21]. There are 211 arrays with 22,277 probesets. We used the ImmGen RMA normalized data and the D-map normalized and batch corrected dataset as previously published[27] [21]. When more than one probeset had the same gene symbol, only the probeset with the highest mean expression was used.

Orthologous gene mapping and filtering

We used Ensembl COMPARA release 63 to map orthologs. Mouse ENSEMBL gene IDs were matched to 15,265 probesets on MoGen and human ENSEMBL gene IDs were matched to 10,457 probesets on U133A. In COMPARA there are 15,678 one-to-one orthologs and 539 apparent one-to-one orthologs resulting in 16,217 one-to-one mapped genes between the species. Of these, 10,248 one-to-one ortholog pairs were measured in both the human and mouse arrays. Finally, only the 5,841 genes with expression level above 120 (recommended ImmGen threshold for expression) in at least three arrays of the common samples (see [23] for details on mapping human-mouse samples, particularly **Supplementary Table 1** there**)** in both species were included in the filtered one-to-one orthologs set. All further analysis was performed on this set.

Conservation of expression (COE)

The COE is a measure of agreement of expression in comparable groups of samples, *i.e.*, lineages, between two species. We previously defined it at the gene level - for each species, we first compute median of the gene's expression in each of the seven common lineages, and then calculate the COE of the gene as the Pearson correlation between these lineages[23]. Similarly, the COE of a module's expression is the COE of the mean expression of the genes in the module (since variation from the module's mean is expected to reflect noise).

Module conservation and comparison

We used the 80 human modules as previously published[21] and 81 mouse modules from the ImmGen Consortium analysis [27]. Of these, we analyzed the 80 human modules and 67 mouse modules containing 5 or more one-to-one orthologs of the filtered set. Modules' conservation was estimated by the $Z_{summary}$ statistic suggested by Langfelder *et al*.[28], using the WGCNA R package. The $Z_{summary}$ statistic estimates the conservation of a reference dataset module in a test set. When testing for conservation of human modules, the full human dataset was used as reference, and the full mouse dataset was used as a test set (**Supplementary Table 1**). Similarly, when testing for conservation of mouse modules, the full mouse dataset was used as reference, and the full human dataset was used as a test set (**Supplementary Table 1**).

We calculated the overlap between each pair of human and mouse modules, and assigned a p-value for module overlap using a hyper-geometric distribution for two samples, with the genes in the filtered set that were assigned to modules in both species serving as background, followed by an FDR of 10%.

Comparison of model regulators

We used the mouse regulatory model defined for the mouse modules by Ontogenet [27], and created a distinct regulatory model for human by applying Ontogenet [27] independently to the human modules. Since the original publication on the human dataset[21] used a different method to create a regulatory model, we avoided comparing directly to it, to avoid artifacts due to difference in analysis tools used in the two species.

As elaborated in [27], Ontogenet has several major advantages over the method originally used in the D-map study (Module Networks[11]), and has been specifically devised to address some of the challenges – and leverage some of the unique power – of studying transcriptional programs in cell lineages. **First**, Ontogenet can identify a whole set of 'equivalent' regulators, whereas the Module Networks approach would have had to (somewhat arbitrarily) choose only one representative. This is more consistent with the dense interconnected nature of regulatory circuits that control cell states. **Second**, Ontogenet allows us to choose a regulator in a context-specific manner, assuming that it may be relevant to the regulation of a gene module only in some cells in the lineage, but not others. **Third**, Ontogenet uses the lineage tree to guide its search for a regulatory program, by preferring (but not mandating) models where 'close' cells in the lineage share regulatory mechanisms.

The candidate regulators for the mouse model (from which the method chooses active regulators) followed those previously reported for human[21], with minor adjustments described in [27]. We then used their one-to-one orthologs as candidate regulators for the human model

(**Supplementary Table 3**). The targets of a regulator were defined as the union of all modules to which it was assigned. Significant overlaps between targets of the 155 regulators chosen by Ontogenet in the two species were tested for using the hyper-geometric test with an FDR of 10%. The background was all genes assigned to modules in both species and included in the filtered one-to-one orthologs set. To test if the number of orthologous regulator pairs whose targets significantly overlap (13-out-of-155) is higher than would be expected by chance, we permuted the regulatory interactions of the 155 common regulators in one species while preserving the regulatory interactions at the module level and repeated the calculation 1,000 times. When comparing regulators of significantly overlapping (orthologous) modules, we used a hyper-geometric test to test the significance of the overlap with FDR of 10%, using as background all 155 candidate regulators that are included in the filtered one-to-one orthologs set and were chosen by Ontogenet.

Ontogenet also provides for each lineage a list of activators and repressors, based on their average regulatory weight across all cell types in the lineage and all modules. For each lineage, we used a hyper-geometric test with FDR 10% to test the significance of the overlap of the lineage activators in human and mouse, and similarly of the lineage repressors in the two species. The background was all regulators that were chosen by the model in either species and are included in the filtered one-to-one orthologs set. FDR was applied to the p-values of all 14 regulators groups (7 lineages, activators and repressors for each).

Motif enrichment in modules

Motif scanning and Motif scoring threshold were done as previously described[27], resulting in a MAX-LOD(i,k) score for each motif k in each gene i, reflecting the best motif instance over the entire promoter region. 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 in that species 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 in that species 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 pairs of modules and motifs that were assigned p-values passing this procedure.

Estimating the effect of conserved *cis*-regulatory elements on expression

We used a one sided KS-test to estimate if the 286 human and mouse module pairs whose genes are enriched for the same motif are more similar in expression pattern, as defined by a COE measure. The background was a COE distribution generated from 1,000 repeats of the COE of 286 randomly selected modules from all human modules enriched for any motif and 286 randomly selected modules from all mouse modules enriched for any motif.

Multiple comparison control

The Benjamini Hochberg False Discovery Rate (FDR[38]) procedure was used to control the false discovery rate at 5% or 10%, whenever stated.

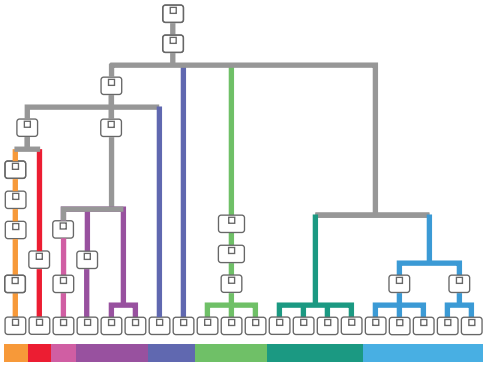
References

1. Shultz LD, Ishikawa F, Greiner DL. Humanized mice in translational biomedical research. Nat Rev Immunol. 2007;7(2):118-30.
2. Gordon J, Grafton G, Wood PM, Larché M, Armitage RJ. Modelling the human immune response: can mice be trusted? Current Opinion in Pharmacology. 2001;1(4):431-5.
3. Payne KJ, Crooks GM. Immune-Cell Lineage Commitment: Translation from Mice to Humans. Immunity. 2007;26(6):674-7.
4. Khanna R, Burrows SR. Human immunology: a case for the ascent of non-furry immunology. Immunol Cell Biol. 2011;89(3):330-1.
5. Davis MM. A prescription for human immunology. Immunity. 2008;29(6):835-8. Epub 2008/12/23. doi: S1074-7613(08)00511-6 [pii]

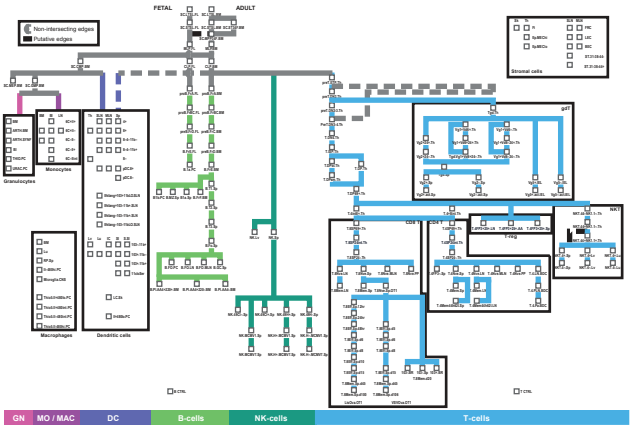
10.1016/j.immuni.2008.12.003. PubMed PMID: 19100694; PubMed Central PMCID: PMC2905652.
6. Rangarajan A, Weinberg RA. Comparative biology of mouse versus human cells: modelling human cancer in mice. Nat Rev Cancer. 2003;3(12):952-9.
7. Mestas J, Hughes CCW. Of Mice and Not Men: Differences between Mouse and Human Immunology. The Journal of Immunology. 2004;172(5):2731-8.
8. von Bernuth H, Picard C, Jin Z, Pankla R, Xiao H, Ku C-L, et al. Pyogenic Bacterial Infections in Humans with MyD88 Deficiency. Science. 2008;321(5889):691-6. doi: 10.1126/science.1158298.
9. Cohen AC, Nadeau KC, Tu W, Hwa V, Dionis K, Bezrodnik L, et al. Cutting Edge: Decreased Accumulation and Regulatory Function of CD4+CD25high T Cells in Human STAT5b Deficiency. The Journal of Immunology. 2006;177(5):2770-4.
10. Boushey HA, Fahy JV. Targeting cytokines in asthma therapy: round one. Lancet. 2000;356(9248):2114-6. Epub 2001/02/24. doi: S0140-6736(00)03486-3 [pii]

10.1016/S0140-6736(00)03486-3. PubMed PMID: 11191533.
11. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet. 2003;34(2):166-76. doi: http://www.nature.com/ng/journal/v34/n2/suppinfo/ng1165_S1.html.
12. Kim HD, Shay T, O'Shea EK, Regev A. Transcriptional Regulatory Circuits: Predicting Numbers from Alphabets. Science. 2009;325(5939):429-32. doi: 10.1126/science.1171347.
13. Rhind N, Chen Z, Yassour M, Thompson DA, Haas BJ, Habib N, et al. Comparative Functional Genomics of the Fission Yeasts. Science. 2011;332(6032):930-6. doi: 10.1126/science.1203357.
14. Tirosh I, Weinberger A, Carmi M, Barkai N. A genetic signature of interspecies variations in gene expression. Nat Genet. 2006;38(7):830-4. doi: http://www.nature.com/ng/journal/v38/n7/suppinfo/ng1819_S1.html.
15. Tirosh I, Wong KH, Barkai N, Struhl K. Extensive divergence of yeast stress responses through transitions between induced and constitutive activation. Proceedings of the National Academy of Sciences. 2011;108(40):16693-8. doi: 10.1073/pnas.1113718108.
16. Barreiro LB, Marioni JC, Blekhman R, Stephens M, Gilad Y. Functional Comparison of Innate Immune Signaling Pathways in Primates. PLoS Genet. 2010;6(12):e1001249.
17. Ravasi T, Suzuki H, Cannistraci CV, Katayama S, Bajic VB, Tan K, et al. An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man. Cell. 2010;140(5):744-52.
18. Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, et al. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. Nat Genet. 2007;39(6):730-2. doi: http://www.nature.com/ng/journal/v39/n6/suppinfo/ng2047_S1.html.

19. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, et al. Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. Science. 2010;328(5981):1036-40. doi: 10.1126/science.1186176.
20. Dermitzakis ET, Clark AG. Evolution of Transcription Factor Binding Sites in Mammalian Gene Regulatory Regions: Conservation and Turnover. Molecular Biology and Evolution. 2002;19(7):1114-21.
21. Novershtern N, Subramanian A, Lawton LN, Mak RH, Haining WN, McConkey ME, et al. Densely Interconnected Transcriptional Circuits Control Cell States in Human Hematopoiesis. Cell. 2011;144(2):296-309.
22. Heng TSP, Painter MW, Elpek K, Lukacs-Kornek V, Mauermann N, Turley SJ, et al. The Immunological Genome Project: networks of gene expression in immune cells. Nat Immunol. 2008;9(10):1091-4.
23. Shay T, Jojic V, Zuk O, Rothamel K, Puyraimond-Zemmour D, Feng T, et al. Conservation and divergence in the transcriptional programs of the human and mouse immune systems. Proc Natl Acad Sci U S A. 2013;110(8):2946-51. doi: 10.1073/pnas.1222738110. PubMed PMID: 23382184; PubMed Central PMCID: PMCPMC3581886.
24. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences. 1998;95(25):14863-8.
25. Stuart JM, Segal E, Koller D, Kim SK. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. Science. 2003;302(5643):249-55. doi: 10.1126/science.1087447.
26. Bergmann S, Ihmels J, Barkai N. Similarities and Differences in Genome-Wide Expression Data of Six Organisms. PLoS Biol. 2003;2(1):e9.
27. Jojic V, Shay T, Sylvia K, Zuk O, Sun X, Kang K, et al. ImmGen Report: Modularity and regulation of gene expression in the mouse immune system. Nat Immunol (in revision). 2013.
28. Langfelder P, Luo R, Oldham MC, Horvath S. Is My Network Module Preserved and Reproducible? PLoS Comput Biol. 2011;7(1):e1001057.
29. Tanay A, Regev A, Shamir R. Conservation and evolvability in regulatory networks: The evolution of ribosomal regulation in yeast. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(20):7203-8. doi: 10.1073/pnas.0502521102.
30. Wapinski I, Pfiffner J, French C, Socha A, Thompson DA, Regev A. Gene duplication and the evolution of ribosomal protein gene regulation in yeast. Proceedings of the National Academy of Sciences. 2010;107(12):5505-10. doi: 10.1073/pnas.0911905107.
31. Tuch BB, Galgoczy DJ, Hernday AD, Li H, Johnson AD. The Evolution of Combinatorial Gene Regulation in Fungi. PLoS Biol. 2008;6(2):e38.
32. Wohlbach DJ, Thompson DA, Gasch AP, Regev A. From elements to modules: regulatory evolution in Ascomycota fungi. Current Opinion in Genetics & Development. 2009;19(6):571-8.
33. Orkin SH, Zon LI. Hematopoiesis: An Evolving Paradigm for Stem Cell Biology. Cell. 2008;132(4):631-44.
34. Blais A, Dynlacht BD. E2F-associated chromatin modifiers and cell cycle control. Current Opinion in Cell Biology. 2007;19(6):658-62.
35. MacIsaac KD, Fraenkel E. Practical Strategies for Discovering Regulatory DNA Sequence Motifs. PLoS Comput Biol. 2006;2(4):e36.
36. GuhaThakurta D. Computational identification of transcriptional regulatory elements in DNA sequence. Nucleic Acids Research. 2006;34(12):3585-98. doi: 10.1093/nar/gkl372.
37. Hogues H, Lavoie H, Sellam A, Mangos M, Roemer T, Purisima E, et al. Transcription Factor Substitution during the Evolution of Fungal Ribosome Regulation. Molecular Cell. 2008;29(5):552-62.
38. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;57(1):289-300.



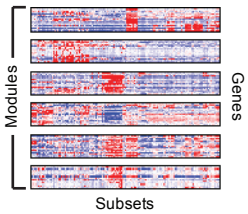
38 cell types
211 samples



249 cell types
816 samples

Expression measurement

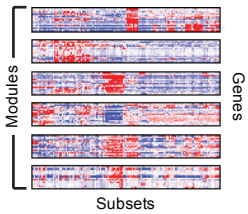
• Gene modules



81 modules

Project modules
Find overlapping modules

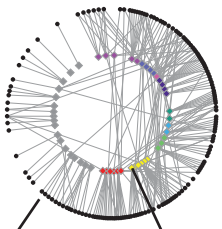
• Gene modules



80 modules

Ontogenet

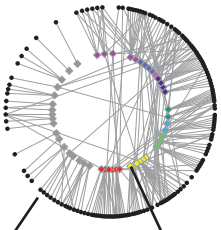
• Regulatory network



Regulators
Modules

Compare regulatory models

• Regulatory network



Regulators
Modules

Figure 1

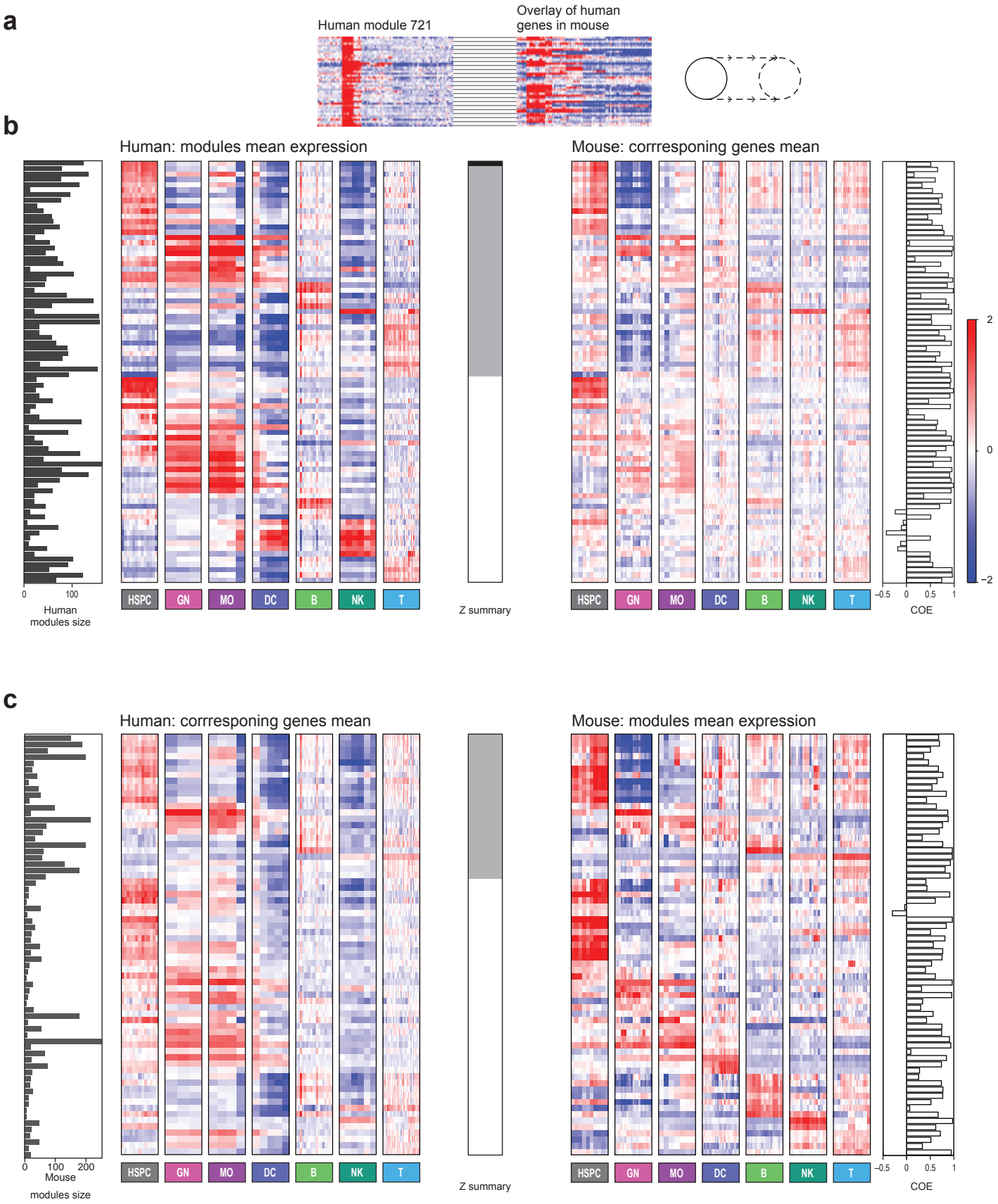

a
Human module 721
Overlay of human genes in mouse
b
Human: modules mean expression
Mouse: corrresponing genes mean
0
100
Human modules size
HSPC
GN
MO
DC
B
NK
T
Z summary
2
0
−2
−0.5
0
0.5
1
COE
c
Human: corrresponing genes mean
Mouse: modules mean expression
0
100
200
Mouse modules size

Figure 2

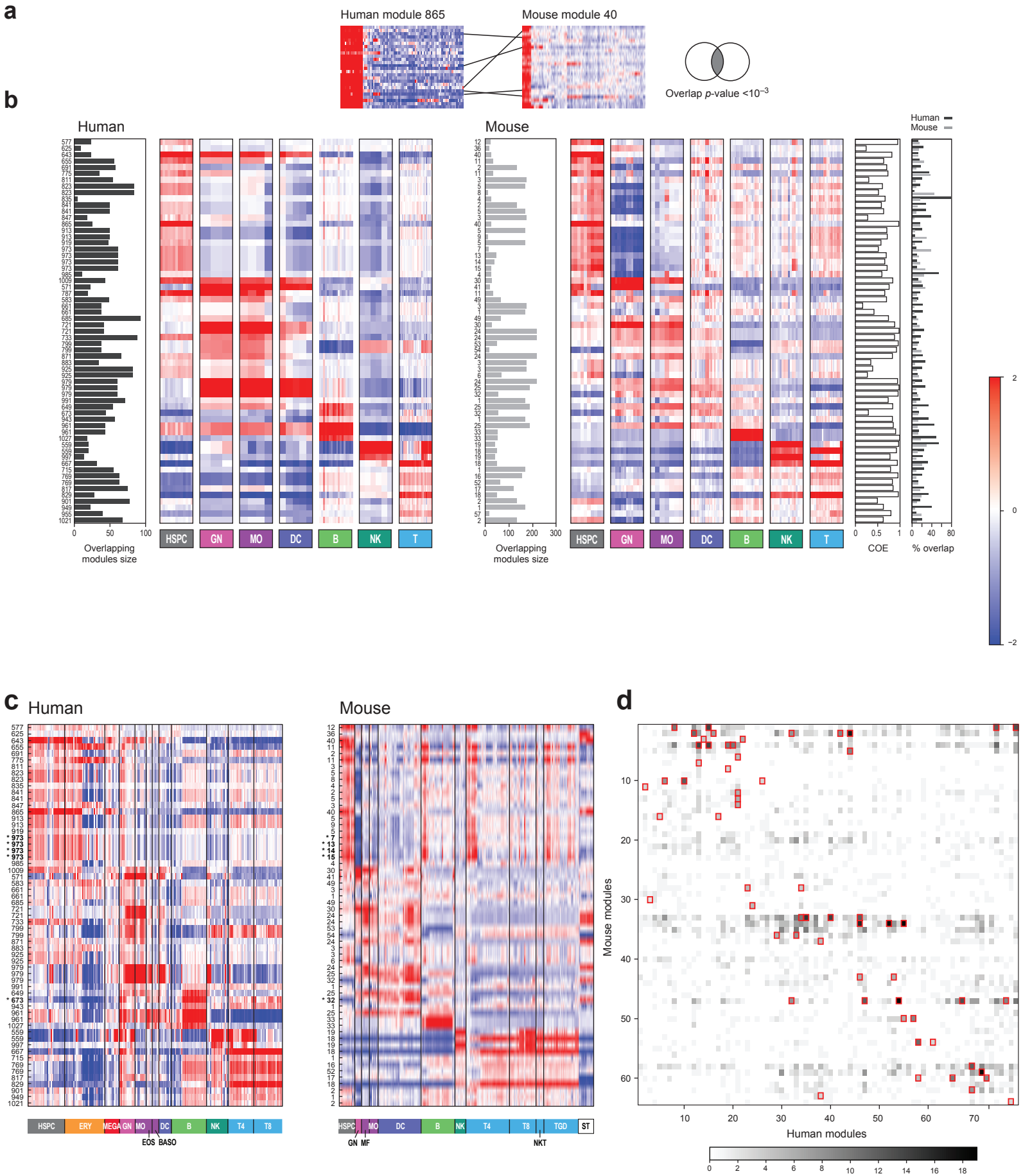

a
Human module 865
Mouse module 40
Overlap p-value <10−3
b
Human
Mouse
Overlapping modules size
HSPC
GN
MO
DC
B
NK
T
COE
% overlap
Human
Mouse
c
Human
Mouse
HSPC
ERY
MEGA
GN
MO
DC
B
NK
T4
T8
EOS
BASO
MF
TGD
ST
NKT
d
Mouse modules
Human modules

Figure 3

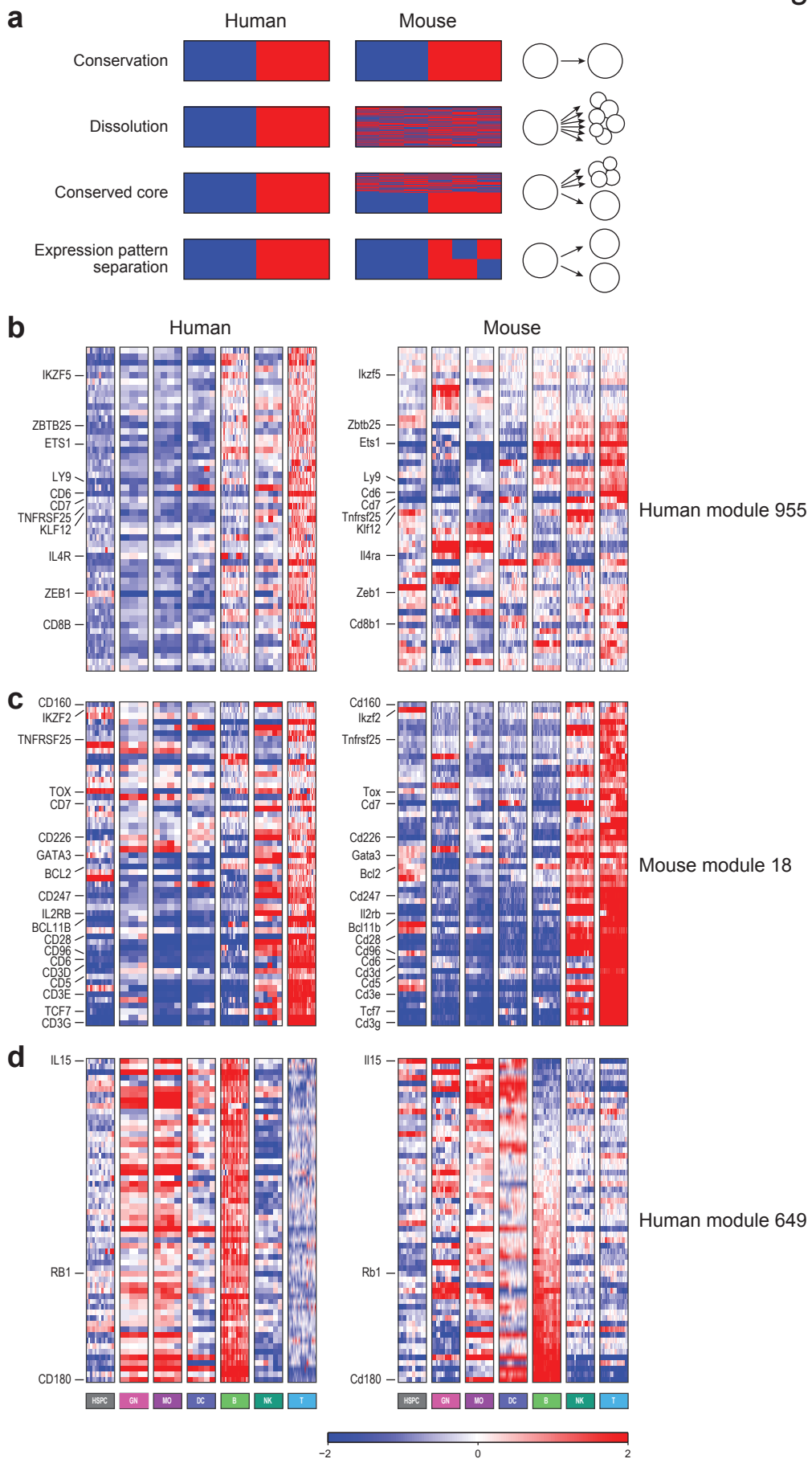
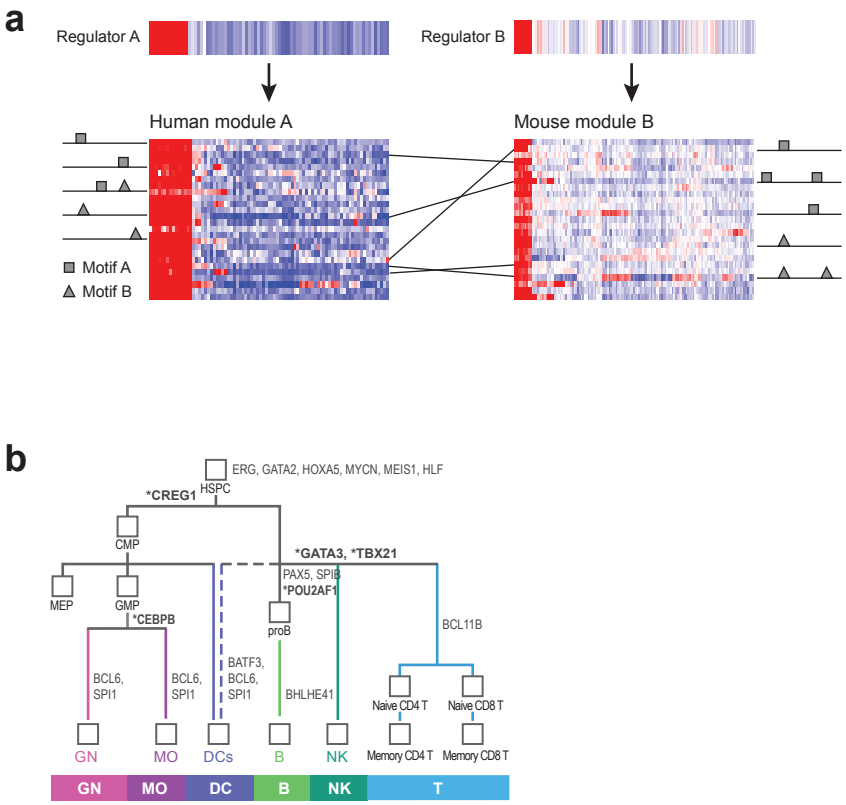
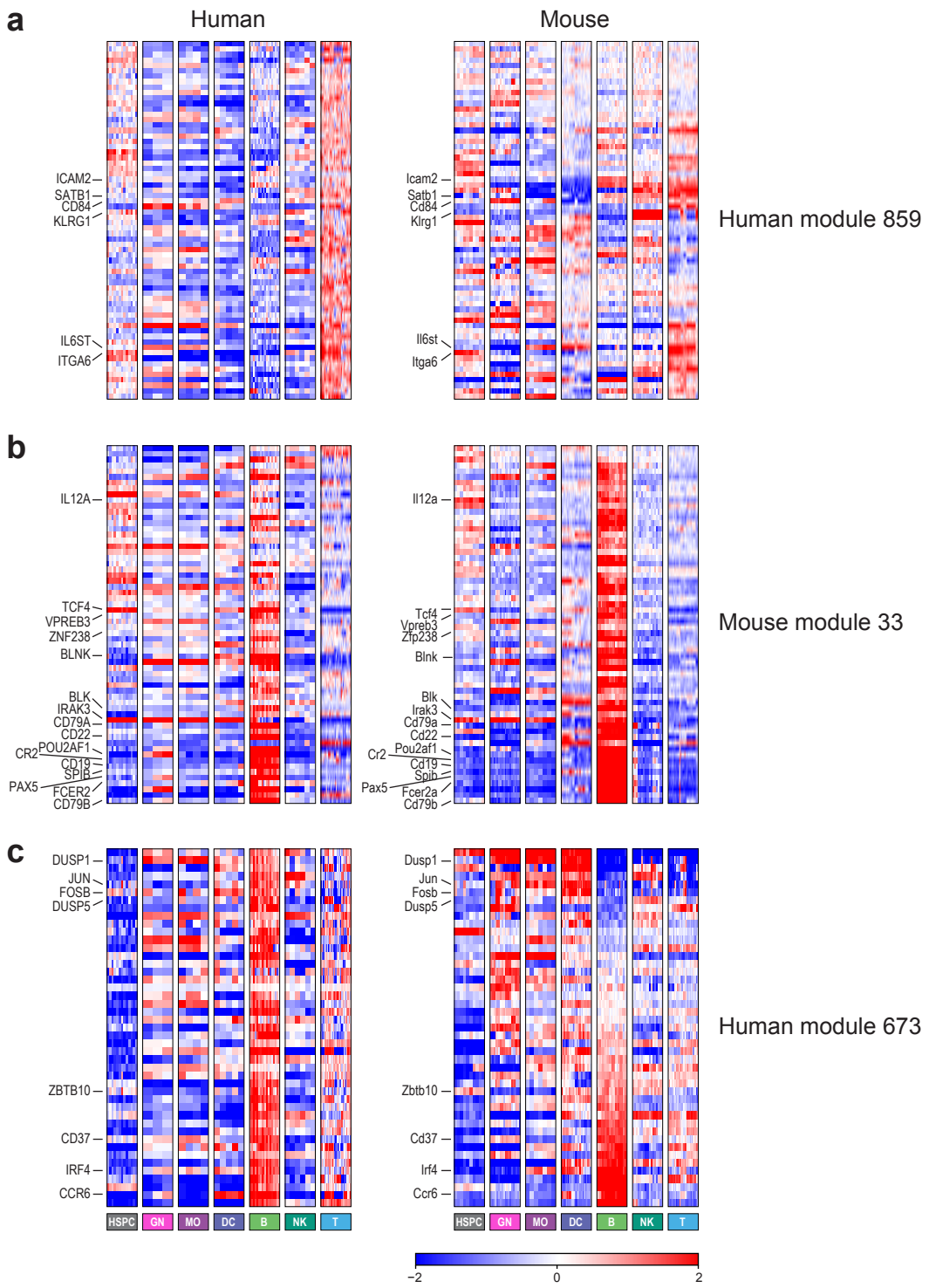

a
Human
Mouse
Conservation
Dissolution
Conserved core
Expression pattern separation
b
Human
Mouse
IKZF5
ZBTB25
ETS1
LY9
CD6
CD7
TNFRSF25
KLF12
IL4R
ZEB1
CD8B
Ikzf5
Zbtb25
Ets1
Ly9
Cd6
Cd7
Tnfrsf25
Klf12
Il4ra
Zeb1
Cd8b1
Human module 955
c
CD160
IKZF2
TNFRSF25
TOX
CD7
CD226
GATA3
BCL2
CD247
IL2RB
BCL11B
CD28
CD96
CD6
CD3D
CD5
CD3E
TCF7
CD3G
Cd160
Ikzf2
Tnfrsf25
Tox
Cd7
Cd226
Gata3
Bcl2
Cd247
Il2rb
Bcl11b
Cd28
Cd96
Cd6
Cd3d
Cd5
Cd3e
Tcf7
Cd3g
Mouse module 18
d
IL15
RB1
CD180
Il15
Rb1
Cd180
Human module 649
HSPC
GN
MO
DC
B
NK
T
−2
0
2

Figure 4


a
Regulator A
Regulator B
Human module A
Mouse module B
Motif A
Motif B
b
ERG, GATA2, HOXA5, MYCN, MEIS1, HLF
HSPC
*CREG1
CMP
*GATA3, *TBX21
PAX5, SPIB
*POU2AF1
MEP
GMP
*CEBPB
proB
BCL11B
BCL6,
SPI1
BCL6,
SPI1
BATF3,
BCL6,
SPI1
BHLHE41
Naive CD4 T
Naive CD8 T
GN
MO
DCs
B
NK
Memory CD4 T
Memory CD8 T
GN
MO
DC
B
NK
T

Supp. Figure 1