ARTICLE

OPEN

doi:10.1038/nature13972

# Conservation of trans-acting circuitry during mammalian regulatory evolution

Andrew B. Stergachis[1]*, Shane Neph[1]*, Richard Sandstrom[1], Eric Haugen[1], Alex P. Reynolds[1], Miaohua Zhang[2], Rachel Byron[2], Theresa Canfield[1], Sandra Stelhing-Sun[1], Kristen Lee[1], Robert E. Thurman[1], Shinny Vong[1], Daniel Bates[1], Fidencio Neri[1], Morgan Diegel[1], Erika Giste[1], Douglas Dunn[1], Jeff Vierstra[1], R. Scott Hansen[1,3], Audra K. Johnson[1], Peter J. Sabo[1], Matthew S. Wilken[4], Thomas A. Reh[4], Piper M. Treuting[5], Rajinder Kaul[1,3], Mark Groudine[2,6], M. A. Bender[7,8], Elhanan Borenstein[1,9,10] & John A. Stamatoyannopoulos[1,3]

**The basic body plan and major physiological axes have been highly conserved during mammalian evolution, yet only a small fraction of the human genome sequence appears to be subject to evolutionary constraint. To quantify cis- versus trans-acting contributions to mammalian regulatory evolution, we performed genomic DNase I footprinting of the mouse genome across 25 cell and tissue types, collectively defining ~8.6 million transcription factor (TF) occupancy sites at nucleotide resolution. Here we show that mouse TF footprints conjointly encode a regulatory lexicon that is ~95% similar with that derived from human TF footprints. However, only ~20% of mouse TF footprints have human orthologues. Despite substantial turnover of the cis-regulatory landscape, nearly half of all pairwise regulatory interactions connecting mouse TF genes have been maintained in orthologous human cell types through evolutionary innovation of TF recognition sequences. Furthermore, the higher-level organization of mouse TF-to-TF connections into cellular network architectures is nearly identical with human. Our results indicate that evolutionary selection on mammalian gene regulation is targeted chiefly at the level of trans-regulatory circuitry, enabling and potentiating cis-regulatory plasticity.**

Gene regulation is classically partitioned into cis- and trans-acting compartments, which are in turn integrated to form a regulatory network. The cis compartment comprises DNA elements that encode TF recognition sites, while the trans compartment encompasses hundreds of TF genes and their DNA recognition repertoires. The cross-regulation of TF genes by one another creates a regulatory network that facilitates complex information processing and potentiates robustness at the cellular and higher levels[1].

In metazoan genomes, actuatable TF recognition sites are clustered into compact (~100–300 bp) regulatory DNA regions that give rise to DNase I hypersensitive sites (DHSs) upon TF occupancy in place of a canonical nucleosome[2]. Mice and humans diverged ~90 million years ago[3], and an extensive survey of mouse DHSs indicates that the cis-regulatory DNA compartment has evolved markedly since the last common ancestor[4], generalizing and extending observations from selected TFs assayed by ChIP-seq in one or a few tissues[5,6]. However, given the limited experimental resolution of previous studies, it is currently unknown how dynamic are individual *in vivo* TF recognition sites within broader regulatory regions, or more generally how cis-regulatory dynamics relate to the conservation of the higher-level cellular and physiological features that define mammals. Earlier studies of individual regulatory elements in *Drosophila*[7] and zebrafish[8] indicate a potential for functional conservation without sequence conservation, and the maintenance of regulatory activity with different phenotypic outcomes. However, the generality of these observations and their broader relevance for mammalian evolution is unclear.

Genomic DNase I footprinting enables systematic delineation of TF–DNA interactions at nucleotide resolution and on a global scale[9–11], permitting: (1) the simultaneous interrogation of hundreds of DNA-binding TFs expressed in a given cell type in a single experiment; (2) *de novo* derivation of the cis-regulatory lexicon of an organism; and (3) systematic mapping of TF-to-TF cross-regulatory networks[1,10].

To delineate an expansive set of specific mouse genomic sequence elements contacted by TFs *in vivo*, we performed genomic DNase I footprinting on 25 diverse mouse cell and tissue types (Extended Data Table 1). From an average of 323 million uniquely mapped DNase I cleavages per cell type, we identified an average of ~1 million high-confidence (false discovery rate (FDR) 1%[10,11]) DNase I footprints (6 to 40 base pairs (bp)), and a total of 8.6 million differentially occupied footprints (Fig. 1a and Extended Data Fig. 1a). DNase I footprints were highly reproducible (Extended Data Fig. 1b) and robust to intrinsic DNase I cleavage propensities (Extended Data Fig. 2a).

## Evolutionary turnover of TF footprints

To study the evolution of TF occupancy patterns between mouse and human, we compared mouse DNase I footprint maps with those from 41 diverse human cell types[10,12] by using bi-directional pairwise alignments of the mouse and human genomes[4] to resolve mouse DNase I footprints to the human genome (Fig. 1b). In total, 65% of mouse TF footprint sequences could be localized within the human genome, comparable to the cross-alignment rate of entire ~150-bp DHSs[4] (Fig. 1c). However, whereas 35% of mouse DHSs have human orthologues that are also DNase I hypersensitive in at least one human cell type[4], only 22% of mouse TF footprints have human sequence orthologues that are occupied in any of the human cell types assayed (Fig. 1c). This indicates that the individual DNA elements within DHSs that are directly contacted

[1]Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA. [2]Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA. [3]Department of Medicine, University of Washington, Seattle, Washington 98195, USA. [4]Department of Biological Structure, University of Washington, Seattle, Washington 98195, USA. [5]Department of Comparative Medicine, University of Washington, Seattle, Washington 98195, USA. [6]Division of Radiation Oncology, University of Washington, Seattle, Washington 98195, USA. [7]Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA. [8]Department of Pediatrics, University of Washington, Seattle, Washington 98195, USA. [9]Department of Computer Science and Engineering, University of Washington, Seattle, Washington 98102, USA. [10]Santa Fe Institute, Santa Fe, New Mexico 87501, USA.
*These authors contributed equally to this work.

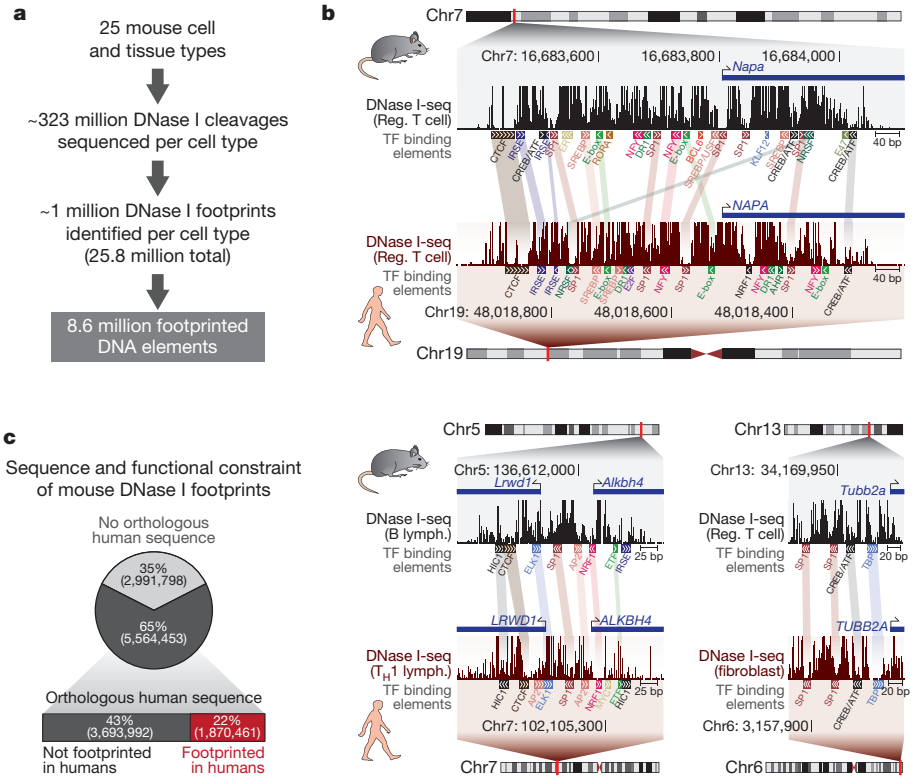



**Figure 1 | Footprinting the mouse genome and comparison with human footprints.** **a**, Derivation of 8.6 million differentially occupied DNase I footprints from 25 mouse cell and tissue types. **b**, Per-nucleotide DNase I cleavage across three gene promoters in both mouse and human cell types; shared TF occupancy sites are indicated by faded boxes. **c**, Percentage of mouse DNase I footprints with sequence aligning to the human genome but not occupied in any human cell type (grey) versus aligning footprints that are occupied in one or more human cell type (red).

by TFs *in vivo* have undergone massive turnover since the last common ancestor of mouse and human.

## Conservation of TF recognition lexicon

Although most mouse TFs have human orthologues, the collective consequences of divergence in DNA binding domains and lineage-specific expansion of certain TF families (for example, KRAB zinc fingers) for the genomic occupancy landscape is unknown. We thus next explored the evolutionary stability of the mammalian TF recognition repertoire encompassed within mouse and human TF footprints. At directly occupied recognition sites for a given TF, footprinting data closely recapitulate TF ChIP-seq[10,11] (Extended Data Fig. 3), and average per-nucleotide DNase I cleavage profiles mirror the morphology of the DNA–protein binding interface[10,11,13]. Examination of cleavage profiles at occupied sites for diverse TFs showed these to be nearly identical between mouse and human cell types (Fig. 2a and Extended Data Fig. 2b), suggesting that *in vivo* DNA recognition preferences for many TFs have experienced little change between mouse and human.

To investigate comprehensively the divergence of mouse and human TF recognition repertoires, we performed *de novo* motif discovery on the 8.6 million mouse TF footprints. In total, we defined 604 unique motif models collectively accounting for the large majority of footprints (Fig. 2b), of which 355 models (59%) matched those within motif databases and 249 were novel (Extended Data Fig. 4a). Comparison of known and novel mouse-derived motif models to motif models derived *de novo* from 8.4 million human DNase I footprints[10] revealed that >94% of the collective TF lexicon is conserved between mouse and humans (Fig. 2c). The human lineage has witnessed expansion of certain TF gene families, notably zinc finger TFs[14]; our results indicate that the proportion of genomic DNA elements bound by lineage-specific TFs *in vivo* is comparatively small. The fact that TF footprints in mouse and human contain highly similar effective *in vivo* recognition sequence repertoires indicates that regulatory divergence between mouse and humans has occurred chiefly at the level of individual TF-binding cis-regulatory elements.

A total of 22 novel motif models were selective for the mouse lineage and 14 were selective for the human lineage (Fig. 2c). The 22 novel mouse-selective motifs are found chiefly in distal elements (Extended Data Fig. 4b), where they populate ~2% of DNase I footprints and show cell/tissue-specific occupancy, predominantly for mouse ES cells (Fig. 2d, e). This suggests that the TFs recognizing these elements may have important roles in very early development, when humans and rodents show more differences than at later stages[15], and further highlights the role of distal gene regulation in species divergence[16]. Notably, whereas sequence matches to the 14 human-selective models in human DNase I footprints showed evidence of strong human-specific evolutionary constraint[10,17] (Fig. 2f), nucleotide diversity at sequence matches to the 22 mouse-selective models in human DNase I footprints is compatible with significantly reduced human-specific evolutionary constraint ($P < 0.05$) (Fig. 2f), consistent with a loss of TF occupancy (and selective pressure) due to divergence (or loss) of the cognate factor within the human lineage.

## Conservation of TF–to–TF connections

We next sought to characterize the core mouse TF regulatory network, and to compare its features with the human TF network. Genomic footprinting provides a direct and empirical approach for mapping the core TF regulatory network of an organism comprising cross-regulatory interactions (network edges) between TF genes (network nodes)[1]. Footprint-anchored TF regulatory networks precisely recapitulate well-validated TF-to-TF regulatory connections[1,18], and are agnostic to whether any given TF-to-TF regulatory interaction is positive (activating) or negative (repressive), as these may vary conditionally even for a given TF. Following the approach of ref. 1, we mapped mouse TF-to-TF networks connecting the 586 mouse TF genes with known recognition sequences (Supplementary Information) within each of the 25 cell/tissue types

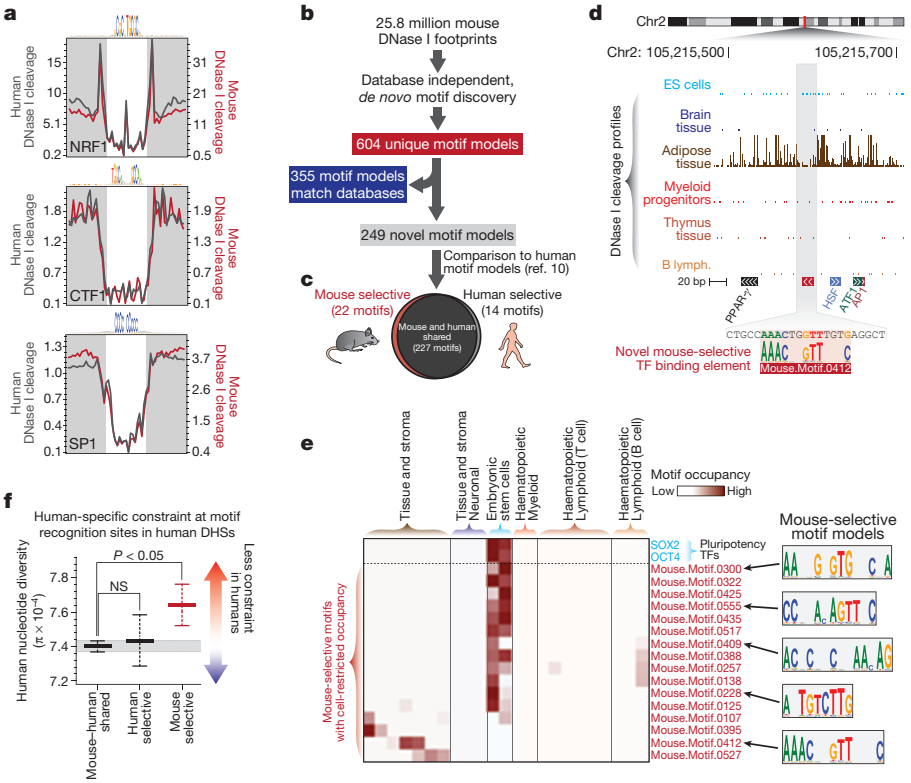



**Figure 2 | Mouse TF footprints define a conserved cis-regulatory lexicon.** **a**, Average per-nucleotide DNase I cleavage at occupied TF recognition sites within mouse and human DHSs. **b**, Of 604 motif models derived *de novo* from mouse footprints, 355 match curated databases. **c**, Comparison of 249 novel mouse motif models with models derived from human footprints. **d**, DNase I footprinting pattern at a novel mouse-selective motif instance. **e**, Preferential occupancy of 16 out of 22 mouse-selective motifs (red); occupancy of pluripotency-related TFs is shown in blue. **f**, Average human nucleotide diversity (π) in different classes of human DNase I footprints partitioned by matches to mouse-derived motifs (mean ± 95% confidence interval (CI); bootstrap resampling). NS, not significant.

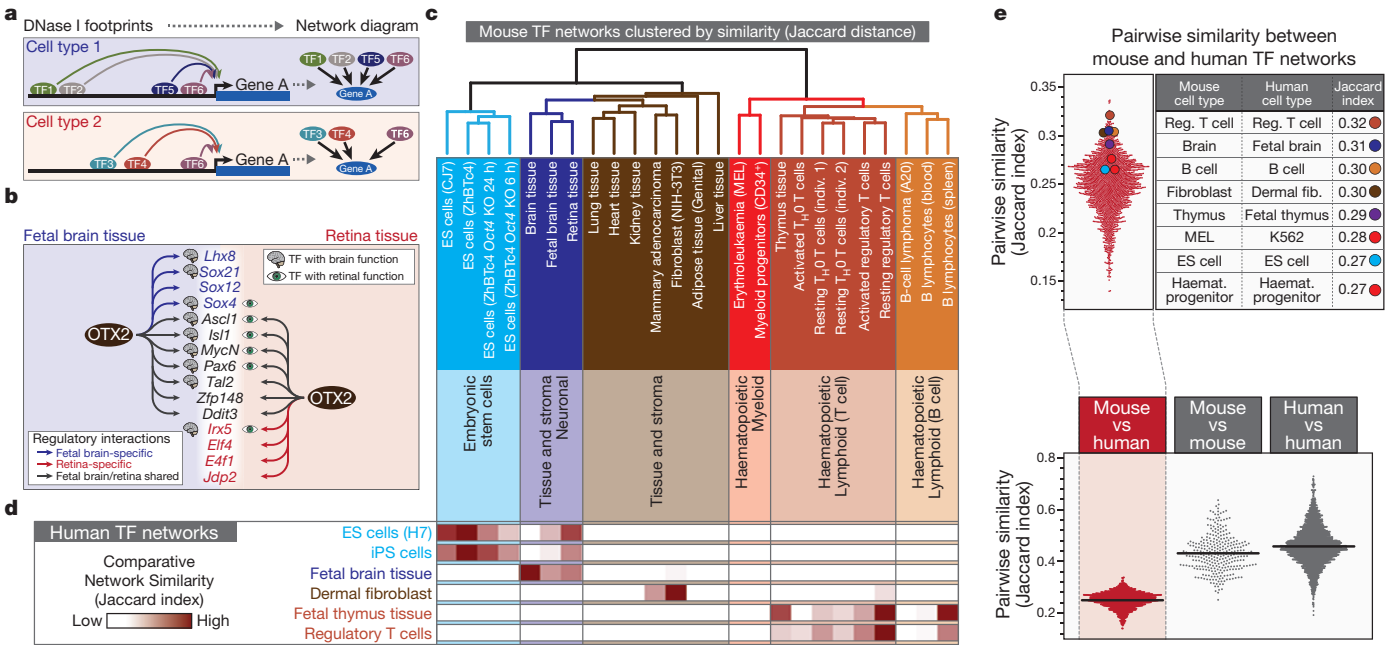



**Figure 3 | Evolutionary dynamics of cis-regulatory logic.** **a**, Schematic for construction of cell-type regulatory networks using TF footprints: TF genes = network nodes; occupied TF motifs = directed network edges. **b**, TF genes regulated by OTX2 in fetal brain and retina networks. Symbols indicate known roles of target genes in brain versus retina development. **c**, Clustering of cell/tissue TF regulatory networks using Jaccard distances between regulatory networks. Cell/tissue types are coloured using physioanatomical and/or functional properties. **d**, Heat map showing network similarity (Jaccard index) between human and mouse cell-type regulatory networks. **e**, Pairwise similarities (Jaccard index) between the regulatory networks of all human and mouse cell/tissue types.

(Fig. 3a). This disclosed an average of 22,970 unique TF-to-TF edges per cell type, totalling 77,084 non-redundant edges across all 25 cell types. Differences between cell types derived from both the cell-selective usage of TFs, as well as the cell-selective occupancy patterns of these TFs. For example, the neuronal developmental regulator OTX2 is selective for neuronal tissue, but its connectivity/occupancy patterns differ between distinct neuronal cell/tissue types (Fig. 3b).

Mouse TF regulatory networks from functionally similar cell and tissue types are coherently organized into anatomical and functional groups (Fig. 3c), analogous to results from human TF regulatory networks[1]. However, although the similarity (pairwise Jaccard indices) between all mouse and human networks was mostly maximal between orthologous mouse–human cell and tissue pairs (Fig. 3d, e), network differences within each species were smaller than differences between species (Fig. 3e).

We next asked to what extent specific mouse TF-to-TF regulatory connections were conserved in human. We first identified TF-to-TF connections that were mouse-specific, human-specific or shared across both orthologous human and mouse cell types (Fig. 4a and Extended Data Table 2). We then differentiated shared regulatory edges (that is, present in both a mouse cell type and its human orthologue) arising from TF occupancy of an orthologous binding element from those shared edges arising from occupancy of non-orthologous sequence within regulatory DNA of the orthologous target gene (Fig. 4a). In the former case, both sequence and circuitry are conserved; in the latter, circuitry only. Overall, ~44% of the TF-to-TF regulatory connections are conserved between orthologous mouse and human cell types ($P < 0.001$) (Fig. 4b). However, >40% of these connections represent edges created by TF binding to a novel sequence element arising since mouse–human divergence (Fig. 4b). As such, conservation of functional regulatory circuitry is considerably greater than indicated by sequence conservation alone.

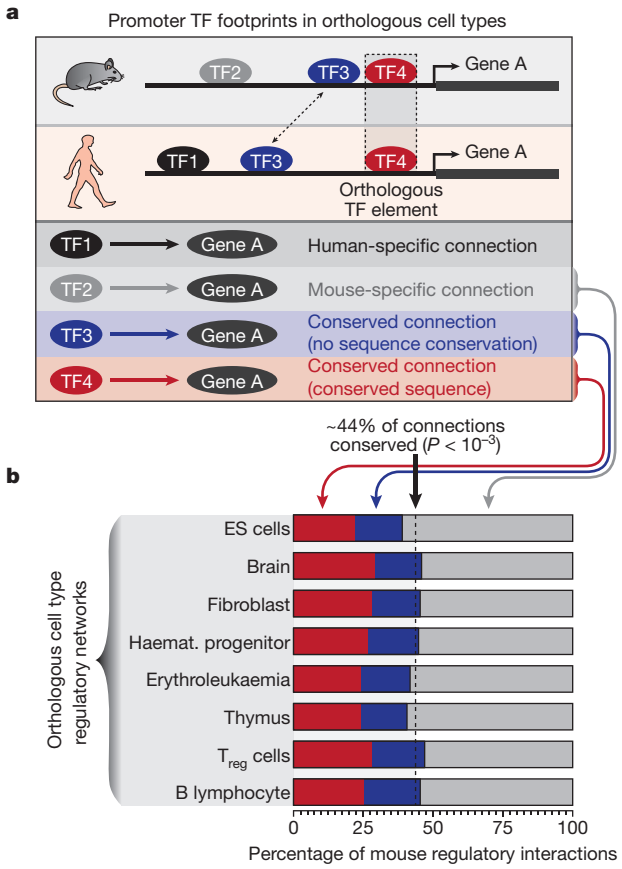



**Figure 4 | Conservation of TF-to-TF regulatory circuitry.** **a**, Four categories of regulatory interactions identified by comparative analysis of mouse and human TF networks. Functionally conserved connections can be mediated by TF occupancy at orthologous (red) or non-orthologous (blue) binding sites. **b**, Categorization and overall conservation of TF-to-TF connections between orthologous mouse and human cell types. On average 44% of TF-to-TF edges are conserved ($P < 0.001$; empirically calculated using shuffled networks).

## Comparative TF network architecture

We next compared the overall architecture of mouse and human TF networks. The architecture of complex networks can be analysed in terms of simple regulatory circuit 'building blocks' termed network motifs, such as the feed-forward loop (FFL)[19]. In human, despite the general selectivity of specific TF-to-TF edges for specific cell types, the pattern of utilization of three-node network motifs within each individual cell type network is nearly identical[1]. Computing network motif utilization within each of the 25 mouse TF networks also revealed uniform patterns across mouse cell/tissue type regulatory networks (Extended Data Fig. 5a). Strikingly, these patterns are nearly identical with human, indicating that mouse and human TF networks utilize virtually the same architecture (Fig. 5a and Extended Data Fig. 5).

To analyse evolutionary conservation at the level of individual regulatory circuits, we identified all instances of each three-node network motif within each mouse cell type, extracted the constituent TFs, and computed how the same TFs were connected in orthologous human cell types. Despite the conservation of overall network architecture between mouse and humans, this analysis revealed that the specific combinations of TFs comprising individual regulatory circuits have undergone substantial remodelling between mouse and human (Fig. 5b and Extended Data Fig. 6). Overall, 39% of combinations of three TFs found within one or more three-node circuit in a given mouse cell type were also organized into at least one type of three-node circuit in an orthologous human cell type (Extended Data Fig. 6b). For example, >25% of three-TF combinations organized into 'regulating mutual' circuits were conserved between orthologous mouse and human cell types, whereas only 8% of three-TF combinations that form 'mutual-and-three-chain' circuits show such conservation. By contrast, 12% of three-TF combinations that form 'mutual-and-three-chain' circuits lose one cross-regulatory interaction, transforming them into FFL circuits in orthologous human cell types (Fig. 5b and Extended Data Fig. 6c). Collectively, TF circuits conserved between mouse and human were enriched in four major network motif types: (1) the FFL motif; (2) the 'regulated mutual' motif; (3) the 'regulating mutual' (RM) motif; and (4) the 'clique' motif (Fig. 5b and Extended Data Fig. 6c). As such, these circuits appear to comprise the most vital building blocks of mammalian TF regulatory architectures.

## Conserved TF positions within networks

We next asked to what degree the position of a specific TF within a given network motif circuit was conserved between mouse and human. To analyse this, we focused on FFL and RM circuits, as these are both strongly conserved overall and have a clear top-down hierarchical organization (Fig. 5a, b). Computation of the propensity for each TF (of 586) to occupy each of the nodes within these network motifs revealed that the preferred position of a given TF within FFL and RM circuits is strongly conserved between orthologous human and mouse cell types (Fig. 5c, d). It also revealed conserved preferential positioning of entire classes of TFs within particular network motif positions. For example, TFs with ubiquitous cellular functions such as CTCF, SP1 and NRF1 systematically localize within the driver positions of FFL and RM circuits (Fig. 5c, d), while TFs involved in cell lineage fate decisions (for example, SOX2, NFE2 and FOXP3) preferentially localized within the final passenger positions (Fig. 5c, d and Extended Data Fig. 7a, b). We also found the passenger edges of FFL and RM motifs to be significantly more cell-selective than the driver edges (Extended Data Fig. 7c, d). These findings raise the possibility that one of the major functions of conserved mammalian network motifs may be to stabilize the expression of TFs that drive cell-type-specific regulatory programs via exploitation of stable cell-ubiquitous regulatory interactions.

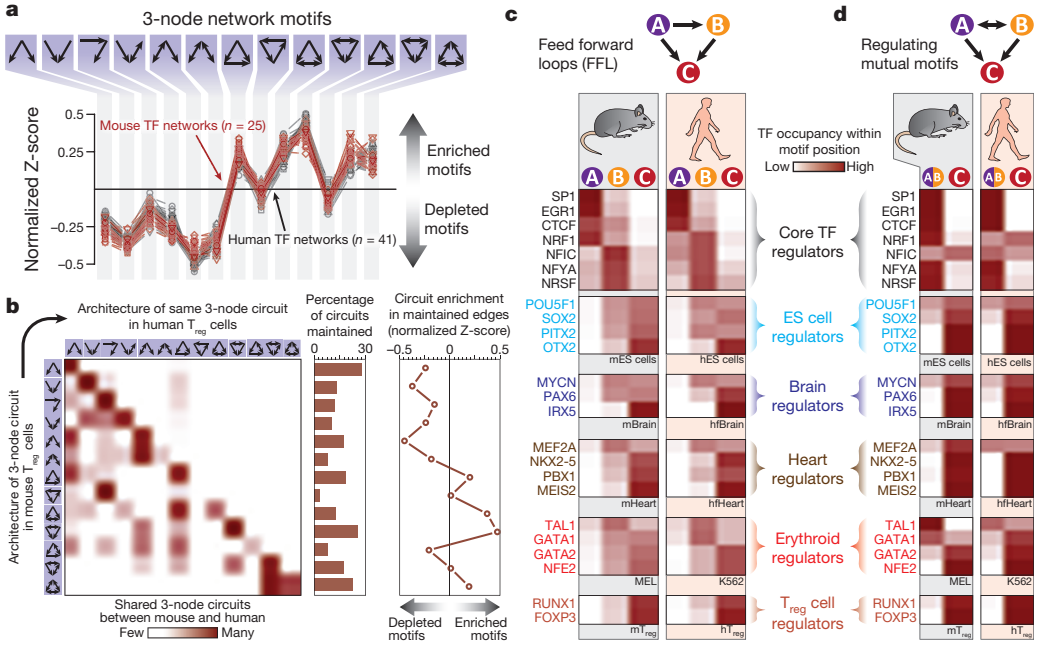



**Figure 5 | Conserved organizing principles of mammalian TF regulatory networks.** **a**, Enrichment of three-node circuits in each mouse (red lines) and human (black lines) TF regulatory network (expanded in Extended Data Fig. 5). **b**, Left: frequency with which individual three-node circuits are identically maintained between the mouse and human $T_{reg}$ network. Middle: percentage of specific three-node circuits identically maintained between the mouse and human $T_{reg}$ network. Right: enrichment of three-node circuits in a network constructed using edges present in both mouse and human $T_{reg}$ networks. **c**, **d**, Frequency with which TFs from six functional classes occupy different positions (driver, first passenger, second passenger) within FFL (**c**) or RM (**d**) circuits in different mouse and human cell-type networks (hfBrain and hfHeart refer to human fetal brain and heart, respectively).

## A conserved developmental program

To explore how the TF regulatory network interacts with downstream non-TF structural/effector genes and to test for conserved interactions, we first quantified, for each TF, whether it preferentially regulates another TF gene(s) or a non-TF 'structural' gene(s) across different mouse and human cell types (Extended Data Fig. 8a). This parameter varied widely between different TFs; in general, TFs involved in development state specification such as HOXB1, OCT4 and SOX2 preferentially regulated other TF genes, while general transcriptional regulators such as NRF1, CTCF and SP1 preferentially regulated non-TF genes (Extended Data Fig. 8b, c). To test how these preferences varied by cell type, we averaged TF gene versus structural gene propensities for all TFs within each cell-type regulatory network. This revealed that the TF networks of pluripotent and early developmental cell types and tissues such as ES cells and fetal brain were globally significantly more oriented towards regulation of TF genes compared with the TF networks of more highly differentiated cell types (for example, B cells, T cells) and tissues (for example, adult brain) (Extended Data Fig. 8d). These TF versus structural gene preferences—both at the individual TF level and at the cell-type regulatory network level—were strongly conserved between mouse and human (Extended Data Fig. 8d, e). The above findings suggest the operation of a conserved global developmental regulatory program that directs a shift in the orientation of TF regulatory networks from TF genes to structural genes during the transition from primitive to definitive cells.

Taken together, our results expose several major organizing principles of mammalian gene regulation, and a fundamental hierarchy in the modes of evolutionary transmission of regulatory information, ranging from poor conservation of cis-acting sequence elements to the preservation of trans-acting and network-level regulatory features (Fig. 6). Conservation of trans-acting components is reflected both in the effective *in vivo* recognition repertoires of human and mouse TFs, which differ only slightly, and in the conserved patterns of TF-to-gene interactions. The dichotomy between cis- and trans-acting regulatory components is most apparent in the context of the core TF regulatory network. Whereas the individual DNA bases contacted by TFs *in vivo* have undergone extensive turnover since the last common ancestor of mouse and human, the repertoire of TFs regulating other TF genes is vastly more conserved. Notably, this cis-acting versus trans-acting disparity in mammals greatly eclipses that previously described for different *Drosophila* species[20].

At the TF network level, organization of the regulatory circuitry in both mouse and human cell types appears to be governed by common principles that result in highly similar network architectures (Fig. 6). Conserved shifts in TF network orientation during the transition from primitive to definitive cells in both organisms suggest that the mammalian regulatory network architecture has converged around a central goal of guiding cell identity during development.

Collectively, our results indicate that evolutionary selection on gene regulation is targeted chiefly at the level of regulatory networks, and

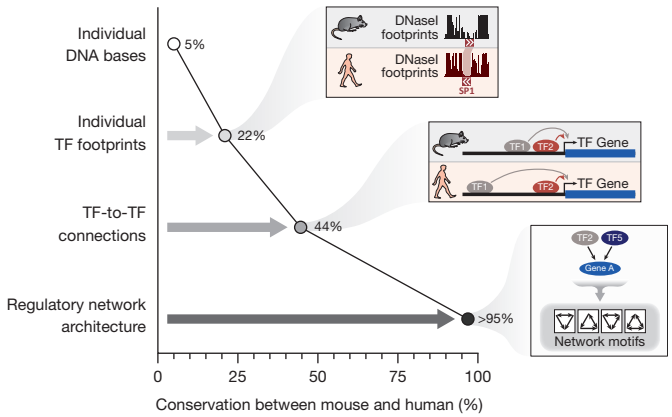



**Figure 6 | Hierarchy of evolutionary constraint on cis- versus trans-regulatory features.** Shown are: overall proportion of conserved DNA bases between mouse and human[3]; proportion of orthologous TF footprints (from data shown in Fig. 1c); average proportion of individual conserved TF-to-TF regulatory connections across orthologous mouse and human cell types (from data shown in Fig. 4); and similarity in overall TF regulatory network architecture (from data shown in Figs 2 and 5).

explain how essential features of the mammalian body plan and physiology have been maintained in the face of massive turnover of the cis-regulatory landscape.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 21 February; accepted 15 October 2014.**


1. Neph, S. *et al.* Circuitry and dynamics of human transcription factor regulatory networks. *Cell* **150,** 1274–1286 (2012).
2. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489,** 75–82 (2012).
3. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
4. Vierstra, J. *et al.* Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science* (in the press).
5. Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328,** 1036–1040 (2010).
6. Villar, D., Flicek, P. & Odom, D. T. Evolution of transcription factor binding in metazoans - mechanisms and functional implications. *Nature Rev. Genet.* **15,** 221–233 (2014).
7. Ludwig, M. Z., Bergman, C., Patel, N. H. & Kreitman, M. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403,** 564–567 (2000).
8. Fisher, S., Grice, E. A., Vinton, R. M., Bessling, S. L. & McCallion, A. S. Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science* **312,** 276–279 (2006).
9. Hesselberth, J. R. *et al.* Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting. *Nature Methods* **6,** 283–289 (2009).
10. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489,** 83–90 (2012).
11. Samstein, R. M. *et al.* Foxp3 exploits a pre-existent enhancer landscape for regulatory T cell lineage specification. *Cell* **151,** 153–166 (2012).
12. Stergachis, A. B. *et al.* Exonic transcription factor binding directs codon choice and affects protein evolution. *Science* **342,** 1367–1372 (2013).
13. Vierstra, J., Wang, H., John, S., Sandstrom, R. & Stamatoyannopoulos, J. A. Coupling transcription factor occupancy to nucleosome architecture with DNase-FLASH. *Nature Methods* **11,** 66–72 (2014).
14. Looman, C., Abrink, M., Mark, C. & Hellman, L. KRAB zinc finger proteins: an analysis of the molecular mechanisms governing their increase in numbers and complexity during evolution. *Mol. Biol. Evol.* **19,** 2118–2130 (2002).
15. Raff, R. A. *The Shape of Life: Genes, Development, and the Evolution of Animal Form* (Univ. Chicago Press, 1996).
16. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188,** 107–116 (1975).
17. Vernot, B. *et al.* Personal and population genomics of human regulatory variation. *Genome Res.* **22,** 1689–1697 (2012).
18. Sullivan, A. M. *et al.* Mapping and dynamics of regulatory DNA and transcription factor networks in *A. thaliana*. *Cell Rep.* **8,** 2015–2030 (2014).
19. Milo, R. *et al.* Network motifs: simple building blocks of complex networks. *Science* **298,** 824–827 (2002).
20. Wittkopp, P. J., Haerum, B. K. & Clark, A. G. Evolutionary changes in cis and trans gene regulation. *Nature* **430,** 85–88 (2004).


**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank our colleagues for their insightful comments and critical readings of the manuscript. We also thank many individuals who provided mouse cell and tissue samples. This work was supported by NIH grants U54HG004592, U54HG007010 and U01ES01156 to J.A.S.; RC2HG005654 to J.A.S. and M.G.; and R37 DK44746 to M.G. and M.A.B. A.B.S. was supported by grant FDK095678A from NIDDK.

**Author Contributions** J.A.S., A.B.S. and S.N. designed the experiments. S.N., A.B.S., A.P.R., E.H. and R.S. carried out the analysis supervised by J.A.S. and E.B.; A.B.S., J.A.S. and S.N. wrote the paper; and all other authors carried out or supervised various aspects of experimental data collection.

**Author Information** All data are available through the mouse ENCODE data repository at UCSC (http://genome.ucsc.edu/ENCODE/) and through GEO series accession GSE51341, or as indicated in Extended Data Table 1. TF regulatory networks may be viewed and downloaded from https://tools.stamlab.org/interactome/mouse and processed data can be downloaded at http://www.mouseencode.org. Human DNase I data can be accessed with GEO series accession GSE51341 and processed data can be viewed and downloaded from http://genome.ucsc.edu/. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.A.S. (jstam@uw.edu).

## METHODS

**Definition of DNase I footprint.** Following the original description of ref. 21, DNase I footprints signify short polynucleotide segments over which the cleavage pattern induced by DNase I is attenuated by the presence of a 'binding protein on the DNA sequence'. This concept was subsequently generalized to encompass altered cleavage patterns encompassing both attenuation of cleavage as well as potentiation of cleavage due to the alteration in minor groove resulting from TF–DNA engagement[22]. It is critical to recognize that DNase I footprints represent TF occupancy at specific positions along the genome. Recently, several publications have mistakenly confounded individual DNase I footprints with aggregated DNase I cleavage profiles for a given TF motif[3,25]. Aggregated DNase I cleavage plots were originated by ref. 9 to visualize and summarize averaged per-nucleotide DNase I cleavage patterns across hundreds to thousands of instances of a given TF recognition sequence (typically within DHSs) genome-wide[9,10]. Because they encompass both occupied and unoccupied motifs, the morphology of the averaged profile depends greatly on the proportion of occupied elements. In the case of TFs with few high-affinity, highly occupied sites, such as the glucocorticoid receptor, aggregated cleavage profiles will dominantly reflect the unoccupied elements, and thus converge on intrinsic DNase I cleavage biases, which have now been well defined[24]. Failure to acknowledge this feature of the data has mistakenly led to erroneous statements concerning DNase I footprinting of low-occupancy TFs, and to restating of previously published conclusions[10,21].

**Genomic footprinting.** A description of each cell and tissue type used in this study can be found in Extended Data Table 1 and at https://genome.ucsc.edu/encode/dataSummaryMouse.html. IACUC approval for all mouse samples was obtained from the Fred Hutchinson Cancer Research Center. Mouse cell and tissue types were subjected to DNase I digestion and high-throughput sequencing, following previous methods[26]. 36-bp sequence tags were aligned to the reference genome, build NCBI37/mm9, using Bowtie 3, version 0.12.7 with parameters: –mm -n 3 -v 3 -k 2, and –phred33-quals. DNase I footprint discovery and false discovery rate estimation (software available at https://github.com/StamLab/footprinting2012) were performed as previously described[10] using 36-mer sequencing reads and unique mappability information for mouse, build NCBI37/mm9 (available at http://www.uwencode.org/proj/hotspot/). For clarity, we note that the footprint detection algorithm we employed differs substantially from (and greatly outperforms) an early algorithm[9]. A recently published modification of the algorithm of ref. 10 termed Wellington incorporates stranded cleavage information and specifically identifies high occupancy sites, although at the expense of greatly reduced sensitivity[27]. Of note, another recently published DNase I footprint detection algorithm[25] was reported to have compared itself against the algorithm of ref. 10, but in fact compared itself against an ad hoc concoction of the ref. 9, ref. 10 and ref. 28 algorithms.

The number and proportion of all DNase I cleavages that fell within DNase I hotspot regions were calculated as previously described[26] (Extended Data Table 1). To identify the total cohort of DNA elements contained within mouse FDR 1% DNase I footprints we first computed the multi-set union of all footprints across all cell types using BEDOPS[29]. For each element of the union, we then collected all significantly overlapping footprints, which were defined as those footprints with 65% or more of their bases in common with the element (bedmap–fraction-map 0.65). A footprint's genomic coordinates were redefined to the minimum and maximum coordinates from its overlap set (bedmap–echo-map-range), which always included the footprint itself. All redefined footprints from the union then passed through a subsumption and uniqueness filter: when a footprint was genomically contained within another, the filter discarded the smaller of the two or selected just one footprint if identical. Footprints passing through the filter comprised the final set of 8.6 million combined footprints across all cell types. Unlike footprints from any single cell type, the combined set included overlapping footprints. We further computed the number of cell types from which each of these 8.6 million combined footprints were derived. To identify the reproducibility of a DNase I footprint, we calculated for every sample the proportion of DNase I footprints that were independently discovered in 1 or more other samples from the same species using an overlap criterion of 25% (bedmap–fraction-either 0.25).

**Accounting for intrinsic DNase I cleavage preferences.** Different rates of DNase I cleavage of phosphate bonds between different flanking base combinations was originally discussed by ref. 21, and have more recently been exhaustively quantified by ref. 24, who performed deep sequencing of DNase I-digested naked DNA from yeast and from human fetal lung fibroblast cells (IMR90) (ref. 24). For each nucleotide $j$ within a genomic window $[i,l]$ the normalized expected cleavage rate is $p_j = a_j / \sum_{k=i}^{l} a_k$. We define $a_k$ as the relative cleavage bias of the 6-mer spanning the positions $[k-3, k+2]$ as described in ref. 24. We redistributed the total observed cleavages ($N_{i,l} = \sum_{k=i}^{l} n_k$) in a window $[i,l]$ such that the observed and expected count for each base $j$ is $n_j$ and $n_j' = N_{i,l} \times a_k$. The per-nucleotide deviation from intrinsic sequence specificity was defined as $\log_2(n_j / n_j')$. The sequence bias normalization was computed separately for each strand and then recombined for visualization purposes.

Using deeply mapped DNase I cleavage preferences[24], we analysed each FDR 1% footprint in all mouse and human cell/tissue types and counted the total number of mapped tags falling in each footprint and the left and right flanking regions. We then randomly assigned the same number of simulated tags to positions within these regions, using probabilities proportional to the DNase I cut-rate bias model for the sequence context surrounding each position. A new footprint-occupancy score (FOS) was calculated over the same L, C and R regions as before[10] and compared to the FOS value of the original footprint. Footprints that showed smaller FOS values using the DNase I cut-rate bias model were considered potential false-positive footprints.

**Correspondence of DNase I footprints with ChIP-seq peaks.** TF occupancy profiles generated by ChIP-seq represent a mixture of both direct (TFs directly contacting the DNA) and indirect (TFs contacting another protein or complex that is contacting the DNA) occupancy events. Of note, for the majority of TFs analysed to date, the indirect component predominates[10]. In contrast to ChIP-seq, DNase I footprinting provides information exclusively at sites of direct TF occupancy[10]. In Extended Data Fig. 3, motif models (from TRANSFAC, JASPAR Core, and UniPROBE) were used in conjunction with the FIMO motif scanning software[30], version 4.6.1 using a $P < 1 \times 10^{-5}$ threshold, to find all motif instances of CTCF (Transfac model V_CTCF_01), GATA1 (Jaspar model MA0035.2-GATA1), MAX (Jaspar model MA0058.1-MAX), Myc (Jaspar model MA0147.1-Myc), and TBP (Transfac model V_TATA_01) within DNase I hotspots of the MEL cell line. We buffered (±30 nucleotides) discovered motif instances and counted at each base position within the buffered motif the number of uniquely mapping DNase I sequencing reads with a 5′ end mapping to that position. We sorted buffered motif instances by their total counts, and then normalized each instance's counts to a mean value of 0 and variance 1. A heat map, with 1 row per motif instance, was generated using matrix2png[31], version 1.2.1. A 46-species phyloP evolutionary conservation score heat map over the same ordered motif instances and bases was generated using the same processing techniques. Motif instances that overlapped DNase I footprints by at least 3 nucleotides were annotated. Uniformly processed mm9 MEL ChIP-seq peaks were downloaded from the UCSC Genome Browser website and motif instances overlapping ChIP-seq peaks by at least 3 nucleotides were also annotated.

**Identification of orthologous human sequence at mouse footprints.** We aligned the coordinates for the 8.6 million combined mouse footprints to the human genome using the 'over chain' best pairwise alignment file available from the UCSC Genome Browser. Mouse footprints with 50% or more of their constituent sequences aligned to the human genome, with at least half not aligned to insertions or deletions, were considered successfully aligned. For a description of the alignment procedure, see ref. 4.

**Aggregated DNase I cleavage profiles.** Mouse motif models from TRANSFAC[32], version 2011.1, JASPAR Core[33], and UniPROBE[34] were used in conjunction with the FIMO motif scanning software, version 4.6.1, using a $P < 1 \times 10^{-5}$ threshold, to find predicted motif instances within hotspot regions as identified by the hotspot algorithm[26]. All motif instances identified for a given model were padded by 10 bp on each side, and aligned in a strand-sensitive manner. DNase I cleavages were averaged for each aligned nucleotide to create an aggregate profile for the motif model.

***De novo* motif model discovery and comparison.** The method for the identification of *de novo* motif models using mouse DNase I footprints was identical to that previously described using human DNase I footprints[10]. Across 25 mouse cell types, we identified 604 unique motif models within DNase I footprints.

We compared *de novo* motif models to models available as part of various experimentally grounded databases, including TRANSFAC, JASPAR Core, and UniPROBE using the TOMTOM software, version 4.6.1 (ref. 35). TOMTOM parameters were set to their default values during model comparisons with the exception of the min-overlap argument, which was set to 5. When partitioning the *de novo* motifs by assigning each to a single category, the order of match assignment preference was to TRANSFAC, JASPAR Core, UniPROBE and finally to the novel motif category. The novel motif models were further classified using previously published motif models derived from human DNase I footprinting experiments[10]. We also determined the proportion of motif models in each experimentally grounded database that matched to mouse *de novo* motif models using TOMTOM with the same parameter settings.

**Analysis of nucleotide diversity (π).** To quantify the nature of selection operating on regulatory DNA, we surveyed nucleotide diversity (π) in DNase I footprints. Population genetics analyses were performed as previously described on 53 unrelated, publicly available human genomes released by Complete Genomics, version 1.10 (ref. 36). Relatedness was determined both by pedigree and with KING[37]. Variant sites were filtered by coverage (>20% of individuals must have calls). Additionally, Complete Genomics makes partial calls at some sites (that is, one allele is A and the

other is N). These were counted as fully missing. Repeats were defined by RepeatMasker, downloaded from the UCSC Genome Browser (http://www.repeatmasker.org). CpGs and repeats were removed from all footprints before analysis. $\pi$ for a single variant is $2pq$, where $p$ = major allele frequency and $q$ = minor allele frequency. $\pi$ was calculated for each cell type by summing for all variants and dividing by total number of bases considered. Although binding elements for mouse-selective motif models are enriched in mouse DNase I footprints, instances of these models in human footprints are also present, but to a significantly lesser degree. To identify instances of mouse-selective motif models in human regulatory elements, human DHSs were scanned using each of the novel mouse-selective motif models and the FIMO software tool ($P < 1 \times 10^{-5}$). Predicted motif instances in human DHSs were then filtered to those that overlapped human DNase I footprints identified in any human cell type by at least three nucleotides.

**Calculation of cell-selective motif occupancy.** We scanned for instances of a motif model using the FIMO software tool ($P < 1 \times 10^{-5}$) and filtered predicted motif instances to those that overlapped DNase I footprints identified in a particular cell type by at least three nucleotides. To derive a final occupancy value for a motif model in that cell type, we counted the total number of DNase I footprinted motif instances for that motif model and normalized it by the total number of bases contained within DNase I footprints in that cell type.

**Calculation of promoter-proximal occupancy of motif models.** We scanned for instances of a novel mouse-selective motif model using the FIMO software tool ($P < 1 \times 10^{-5}$) and filtered predicted motif instances to those that overlapped DNase I footprints identified in any cell type by at least three nucleotides. We classified those within 5 kb of a transcriptional start site using RefSeq annotations as 'promoter-proximal' and all others as 'promoter-distal'.

**TF regulatory network construction.** Transcription factor (TF) regulatory networks were constructed as previously described[1] using 5,000 nucleotide buffers anchored on canonical TF transcriptional start site (TSS) annotations. TF genes and motif models used for network construction were collected from the JASPAR Core, UniPROBE and TRANSFAC databases (Supplementary Information). To create genome-wide networks this method was extended to include all mm9 RefSeq genes, anchored using the 5′-most TSS annotation[38].

**Clustering and similarities of TF regulatory networks.** We computed the pairwise Jaccard distances between TF regulatory networks and applied Ward clustering[39] using the *hclust* and *dendrogram* functions in R. The heat map representation in Fig. 3d used the Jaccard index for a similarity measure. Importantly, all comparisons were made using the same subset of orthologous TF genes (567) with known, associated motif models in both species.

**TF regulatory edge conservation.** To identify conserved regulatory connections that are also sequence conserved we first collected all motif instances that overlapped a DNase I footprint by at least 3 nucleotides in a specific mouse cell type that gave rise to a regulatory edge in that cell-type TF regulatory network. We then aligned the coordinates of this mouse motif instance to the human genome using the 'over chain' best pairwise alignment file available from the UCSC Genome Browser. A mouse motif instance was considered successfully aligned if 50% or more of its underlying sequence aligned to the human genome, with at least half not aligned to insertions or deletions. If a footprinted mouse motif instance aligned to a motif instance of the same TF in an orthologous human cell type that also overlapped a footprint by 3 nucleotides or more, the human motif possibly gave rise to the same regulatory edge. If it did, the edge in the mouse regulatory network was classified as a shared edge between species arising from orthologous binding elements. Notably, an edge that connects two TFs within a regulatory network may arise from a single, or multiple, distinct footprinted TF binding elements. In cases where multiple, distinct footprinted TF binding elements underlie a regulatory edge within a mouse cell-type TF regulatory network, this regulatory edge is considered to arise from an orthologous binding element so long as one of these TF binding elements is a shared connection arising from an orthologous binding element.

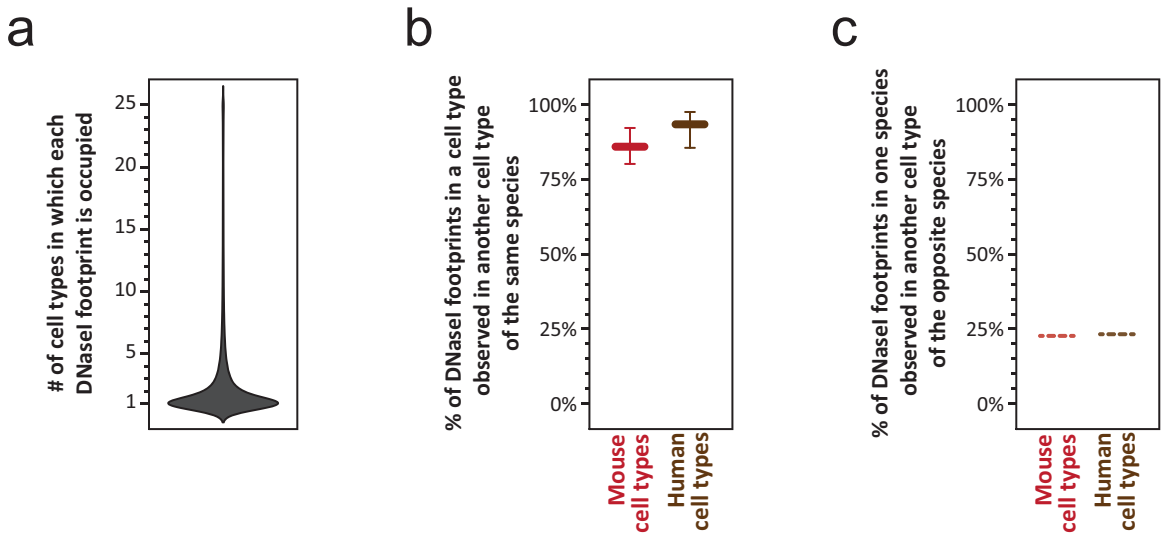
We calculated an empirical *P* value to evaluate the significance of the number of shared edges found between orthologous mouse and human cell types. We first generated 1,000 randomized human TF regulatory networks. When creating a randomized network, we ignored the usual requirement that a motif instance must significantly overlap a human footprint. The genomic space used to construct a random network was identical to that used in the observed case (within 5,000 nucleotides of a canonical TSS). A random subset of generated edges was chosen so that the in-degree to every TF gene node was identical to that of the observed human TF regulatory network case (and, hence, the total number of edges was the same), and all edges were unique. We then determined the number of functionally conserved edges between the observed mouse TF regulatory network and each randomized human TF regulatory network. We counted the number of times this number of functionally conserved edges was at least as large as in the observed TF regulatory network's case. An empirical *P* value was calculated as one more than the number of times this event occurred divided by 1,000. This analysis was performed between every pair of orthologous cell types. No randomized experiment gave a functionally conserved number that reached or exceeded the observed, real TF regulatory networks case.

**Network motif architectures.** We removed self-edges from every TF regulatory network and used the mfinder software tool for network motif analysis[40]. A *z*-score was calculated over each of 13 network motifs of size 3 (three-node network motifs), using 250 randomized networks of the same size for a null estimate. We vectorized z-scores from every cell type and normalized each to unit length to create triad significance profiles[19].

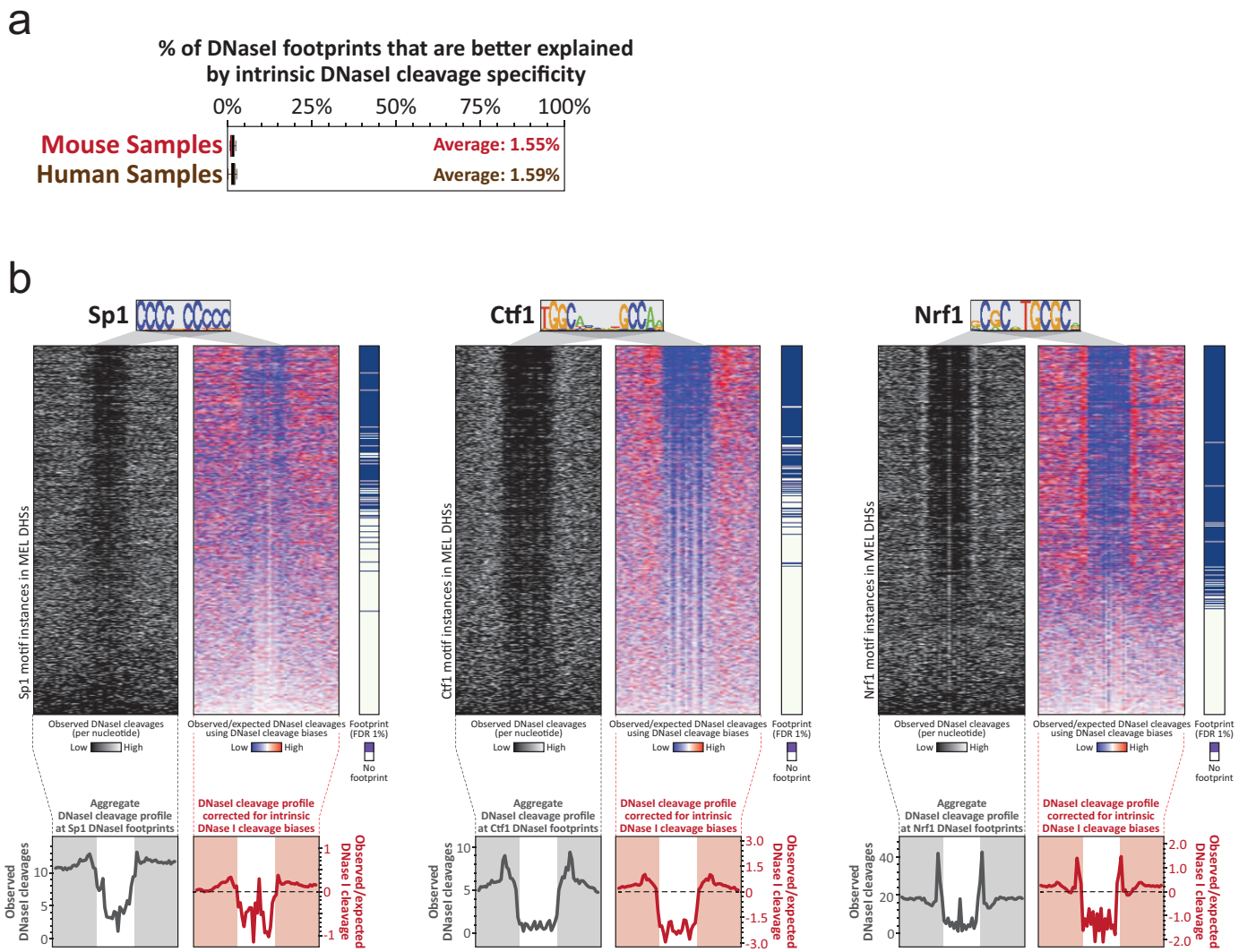
**Distribution of three-node network motifs.** We enumerated all three-node circuits in a mouse TF regulatory network, and determined if and how each was connected in an orthologous human cell-type TF regulatory network. Software is available for download at https://github.com/StamLab/network-motifs.

**Central-facing versus peripheral-facing TF enrichments.** Enrichments were calculated by taking the log base 2 of the ratio of two proportions. The numerator was the proportion of outgoing edges from a TF node in the regulatory network that connected to another TF node, divided by the total number of input edges to all TFs. The denominator was the proportion of outgoing edges from a TF node that connected to any non-TF gene node, divided by the total number of input edges to all non-TFs gene nodes.

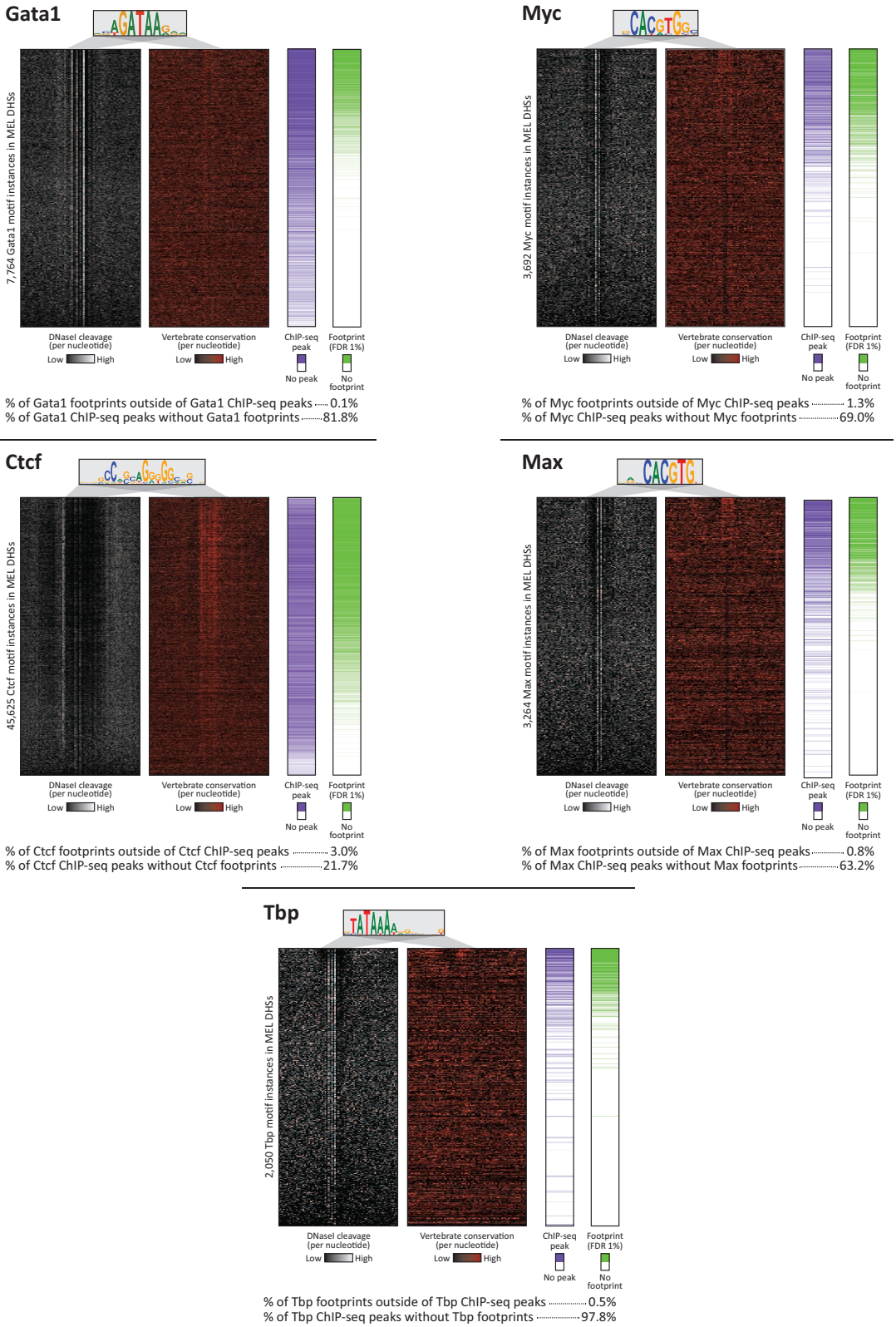

21. Galas, D. J. & Schmitz, A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.* **5,** 3157–3170 (1978).
22. Stamatoyannopoulos, J. A., Goodwin, A., Joyce, T. & Lowrey, C. H. NF-E2 and GATA binding motifs are required for the formation of DNase I hypersensitive site 4 of the human beta-globin locus control region. *EMBO J.* **14,** 106–116 (1995).
23. He, H. H. *et al.* Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nature Methods* **11,** 73–78 (2014).
24. Lazarovici, A. *et al.* Probing DNA shape and methylation state on a genomic scale with DNase I. *Proc. Natl Acad. Sci. USA* (2013).
25. Sung, M. H., Guertin, M. J., Baek, S. & Hager, G. L. DNase footprint signatures are Ddictated by factor dynamics and DNA sequence. *Mol. Cell* http://dx.doi.org/10.1016/j.molcel.2014.08.016 (2014).
26. John, S. *et al.* Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature Genet.* **43,** 264–268 (2011).
27. Piper, J. *et al.* Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res.* **41,** e201 (2013).
28. Mercer, T. R. *et al.* The human mitochondrial transcriptome. *Cell* **146,** 645–658 (2011).
29. Neph, S. *et al.* BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28,** 1919–1920 (2012).
30. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37,** W202–W208 (2009).
31. Pavlidis, P. & Noble, W. S. Matrix2png: a utility for visualizing matrix data. *Bioinformatics* **19,** 295–296 (2003).
32. Wingender, E., Dietze, P., Karas, H. & Knüppel, R. TRANSFAC: A database on transcription factors and their DNA binding sites. *Nucleic Acids Res.* **24,** 238–241 (1996).
33. Bryne, J. C. *et al.* JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.* **36,** D102–D106 (2008).
34. Newburger, D. E. & Bulyk, M. L. UniPROBE: an online database of protein binding microarray data on protein–DNA interactions. *Nucleic Acids Res.* **37,** D77–D82 (2009).
35. Gupta, S., Stamatoyannopoulos, J., Bailey, T. & Noble, W. Quantifying similarity between motifs. *Genome Biol.* **8,** R24 (2007).
36. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327,** 78–81 (2010).
37. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26,** 2867–2873 (2010).
38. Pruitt, K. D. *et al.* The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.* **19,** 1316–1323 (2009).
39. Ward, J. H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **58,** 236 (1963).
40. Milo, R. *et al.* Superfamilies of evolved and designed networks. *Science* **303,** 1538–1542 (2004).



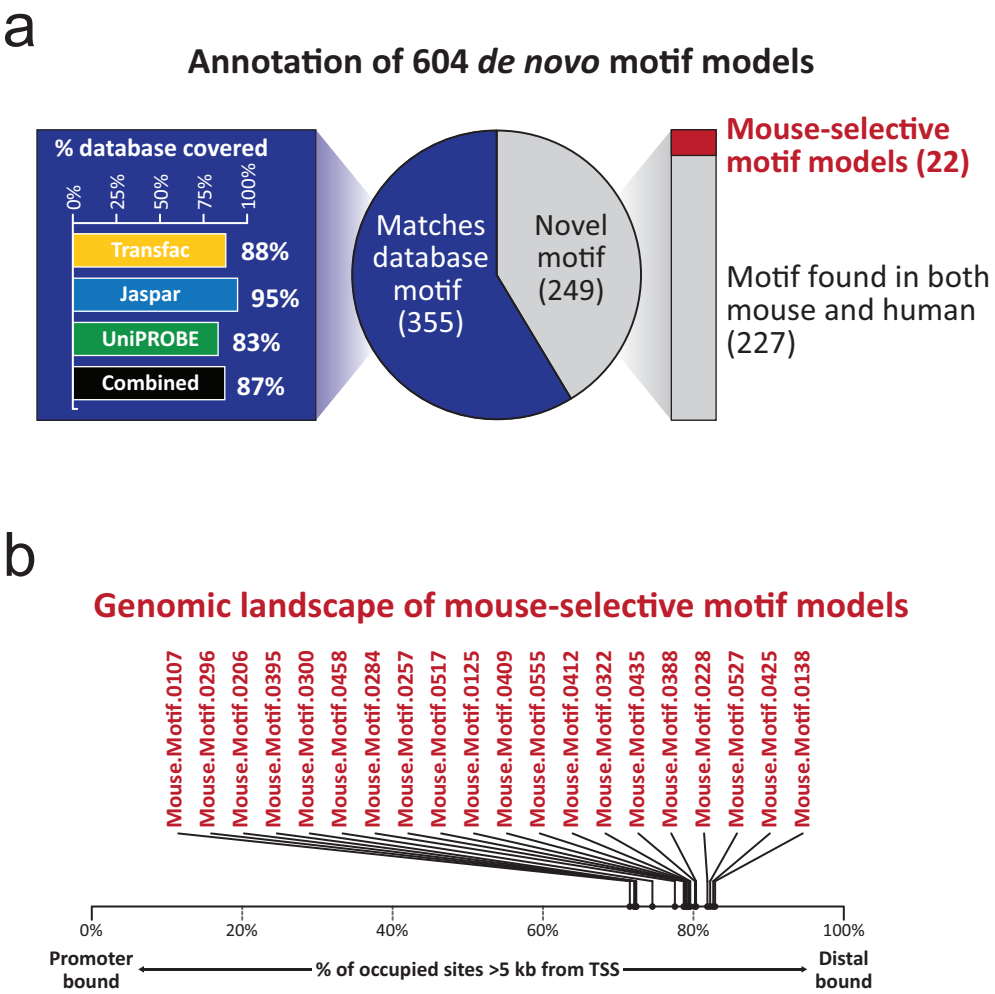

**Extended Data Figure 1 | Cell-selectivity and reproducible detection of DNase I footprints.** **a**, Distribution of the number of mouse cell types in which each of the 8.6 million distinct footprinted cis-regulatory elements in mouse is contained within a DNase I footprint. **b**, For each mouse and human cell type, shown is the percentage of DNase I footprints identified in that cell type that are observed in at least one other mouse or human cell type respectively (data represents median ± 25% and 75% quartiles). **c**, Red: percentage of mouse DNase I footprints with sequence aligning to the human genome that are occupied in one or more human cell types. Brown: percentage of human DNase I footprints with sequence aligning to the mouse genome that are occupied in one or more mouse cell types.



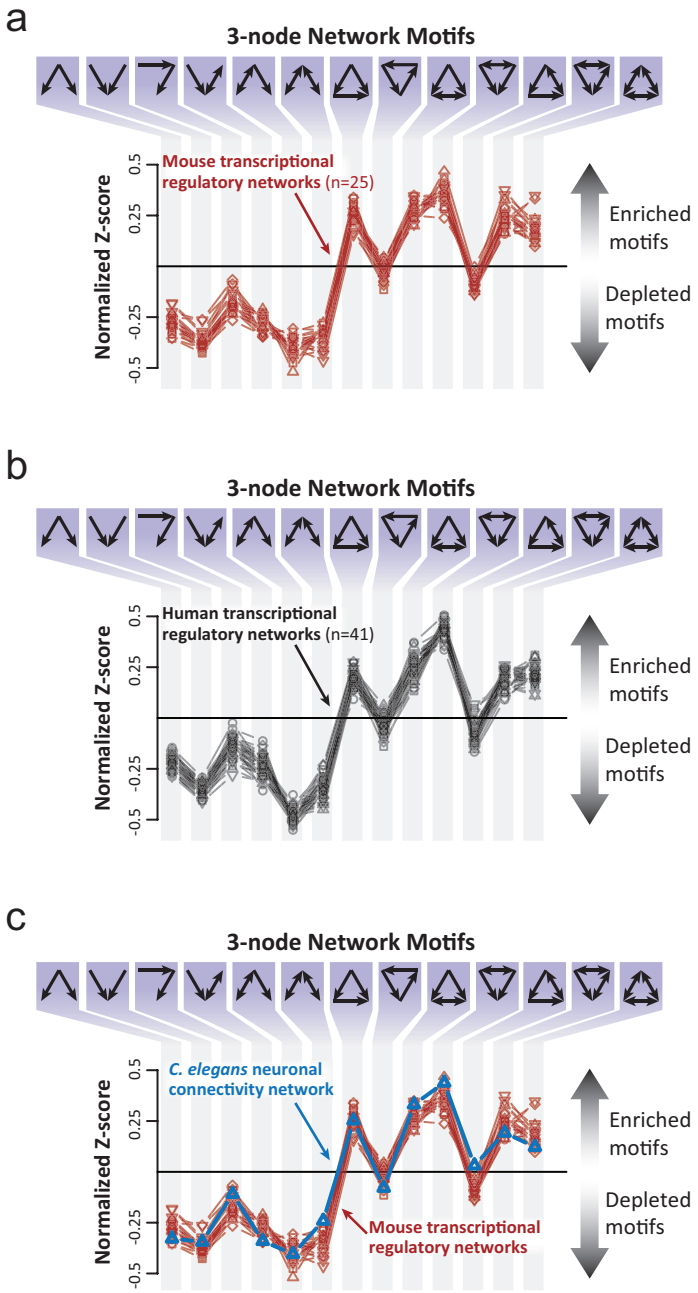

**Extended Data Figure 2 | Negligible impact of intrinsic DNase I cleavage biases on delineation of DNase I footprints.** **a,** Box-and-whisker plot displaying the percentage of DNase I footprints found in each of the mouse and human samples that are potentially better explained by intrinsic DNase I cleavage specificity (box represents mean ± 25% and 75% quartiles and whiskers represent minimum and maximum values across all human and mouse samples, respectively. **b**, Effects of protein occupancy and sequence context on DNase I cleavage profiles. Top: heat maps of per-nucleotide DNase I cleavages; the ratio of the observed cleavages to expected cleavages computed using empirically-modelled DNase I cleavage bias[24]; and discovered 1% FDR DNase I footprints surrounding Sp1, Ctf1 and Nrf1 recognition sequences in MEL cells. Each heat map pixel row corresponds to an individual motif instance within a DNase I hotspot. Each blue tick mark under the 'footprint' column denotes whether (tick) or not (blank) that motif instance overlaps a called FDR 1% DNase I footprint. Bottom: aggregated DNase I cleavage profiles of occupied (that is, within DNase I footprints) Sp1, Ctf1 and Nrf1 recognition sequences in MEL cells shown side-by-side with $\log_2$ ratio of observed versus expected (from intrinsic cleavage preferences) DNase I cleavage. Note that in all cases the cleavage profile of occupied elements differs markedly from expectation.



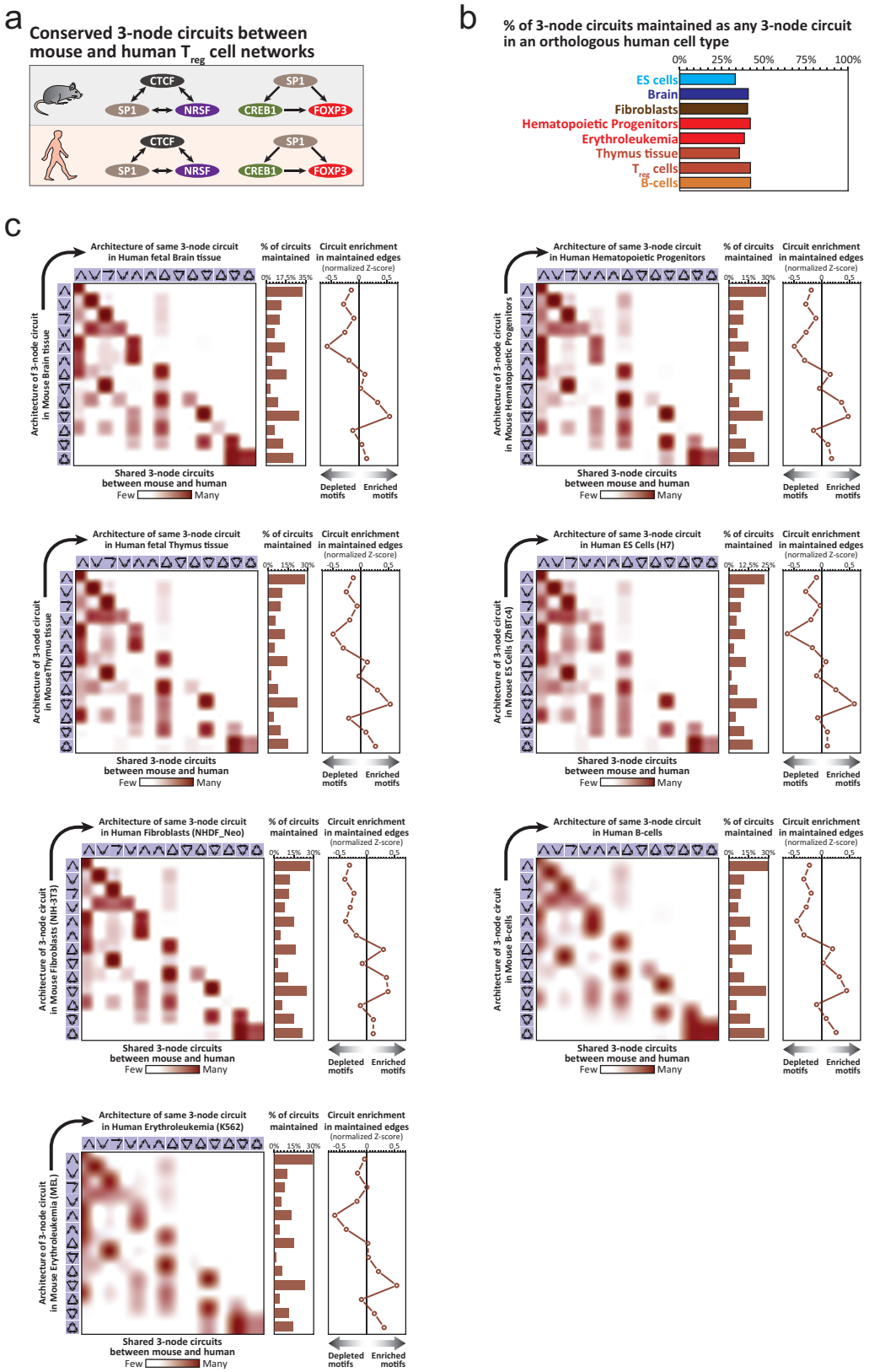

**Extended Data Figure 3 | DNase I footprints accurately recapitulate ChIP-seq data.** For five different TFs with corresponding ChIP-seq data in MEL cells, displayed are (left) heat maps showing per-nucleotide DNase I cleavage and (right) vertebrate conservation by phyloP for all motif instances of that TF within MEL DNase I hotspots (irrespective of whether they overlap a DNase I footprint), ranked by the local density of DNase I cleavages. The number of motif instances for that TF is indicated to the left of the heat map. Purple ticks indicate the presence of the corresponding TF ChIP-seq peaks at each motif instance. Green ticks indicate the presence of DNase I footprints at each motif instance. Below each graph is indicated the percentage of TF footprints that reside outside of a ChIP-seq verified binding site, as well as the percentage of ChIP-seq peaks that do not contain a DNase I footprint for that TF (indicating indirect TF occupancy). Of note, occupied motifs within DNase I footprints accurately recapitulate sites of direct TF occupancy, as 99% of DNase I footprinted motifs for a given TF overlap a cognate ChIP-seq peak. In contrast, for most TFs the majority of ChIP-seq peaks arise from indirect TF occupancy events (and thus lack DNase I footprinted sequence elements for their cognate TF).



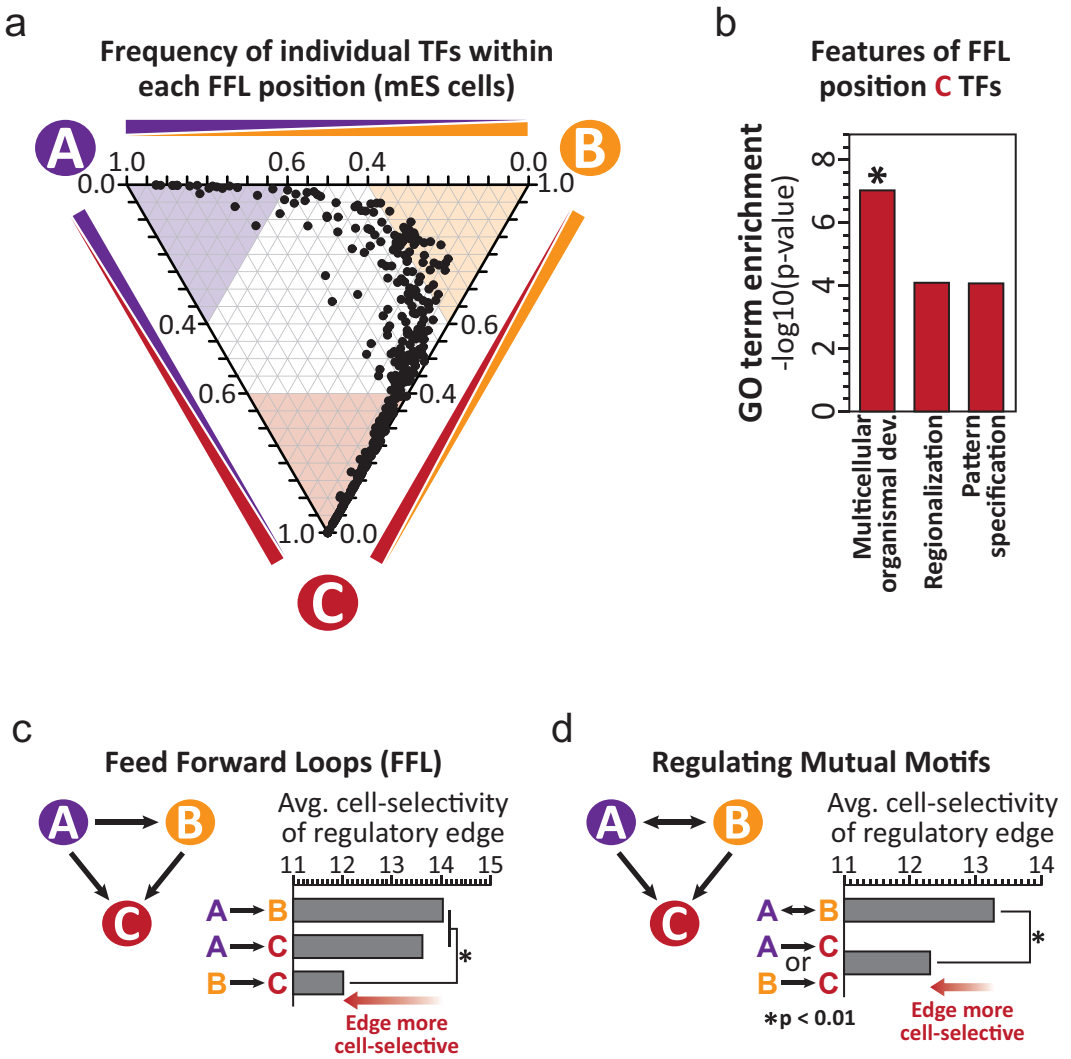

**Extended Data Figure 4 | Annotation of the *de novo* mouse motif models.** **a**, Left: bar chart showing the percentage of the motif models within different experimentally grounded motif databases that match our *de novo* mouse motif models. Right: bar chart showing the number of novel *de novo* motif models in mouse that match *de novo* motif models in human. **b**, The proportion of mouse-selective motif model DNase I footprints within distal regulatory regions.



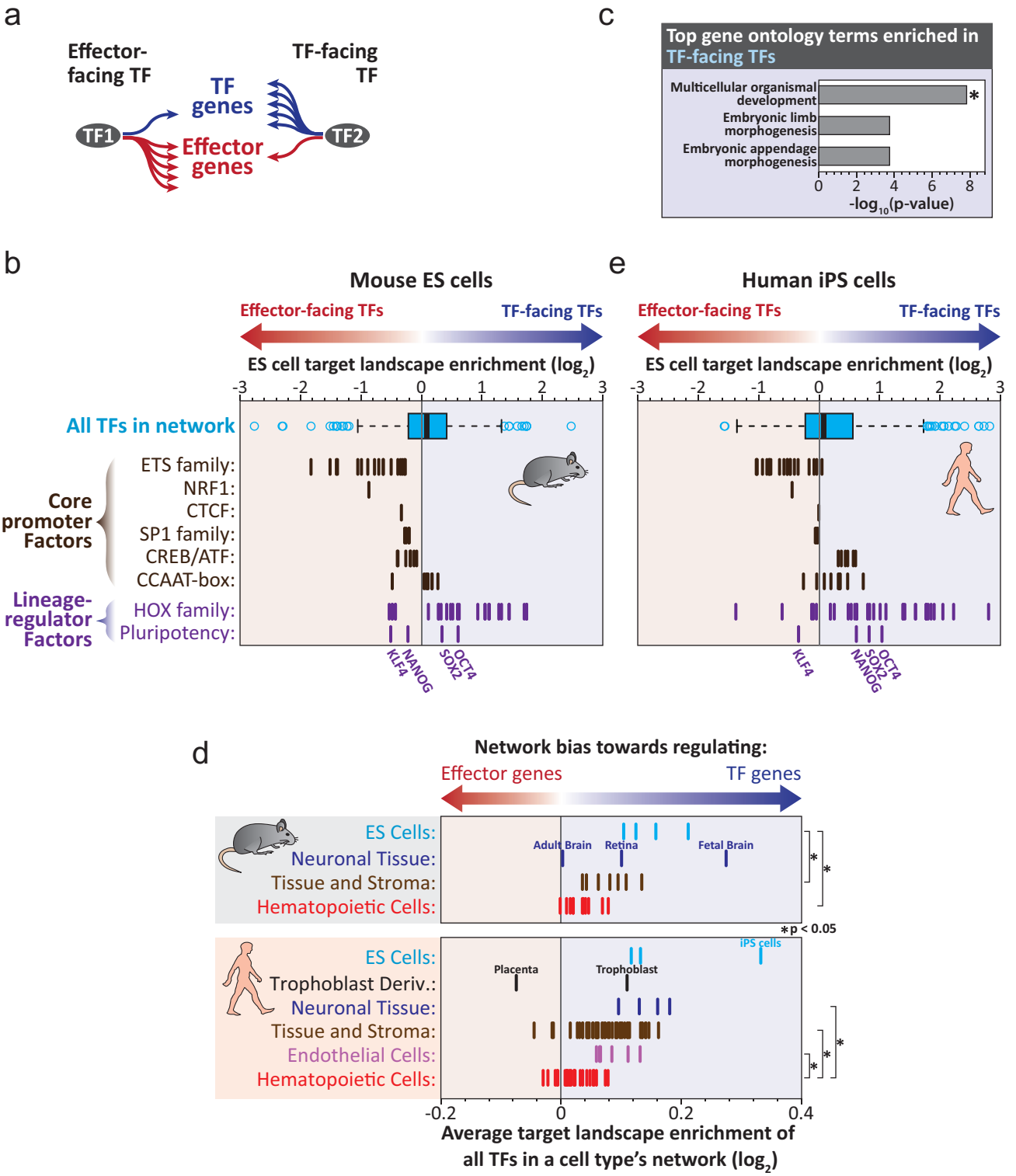

**Extended Data Figure 5 | Conserved organizing principles of the mammalian TF regulatory network.** **a**, **b**, Shown is the relative enrichment or depletion of the 13 three-node network motifs in each of the mouse (**a**) and human (**b**) regulatory networks. **c**, Shown is the relative enrichment or depletion of the 13 three-node network motifs in each of the mouse regulatory networks compared with the relative enrichment of the same motifs in the *C. elegans* neuronal connectivity network.




**Extended Data Figure 6 | The conservation of individual three-node circuit types.** **a**, Examples of three-node circuits formed by TFs in both mouse and human regulatory T ($T_{reg}$) cells. **b**, For each of eight orthologous mouse and human cell-type pairings shown is the percentage of three-node circuits in the mouse cell type that are maintained as any three-node circuit in the orthologous human cell type. **c**, For each of seven orthologous mouse and human cell-type pairings shown is: (left) heat map showing the overall propensity of individual three-node circuits in the mouse cell-type regulatory network to form the same or other three-node circuits in the human cell-type regulatory network; (middle) bar plot showing the percentage of specific three-node circuits in the mouse cell-type regulatory network to be maintained as the same three-node circuits in the human cell-type regulatory network; (right) the relative enrichment or depletion of the 13 three-node network motifs in a regulatory network constructed using the subset of edges present in both mouse and human cell-type regulatory networks.




**Extended Data Figure 7 | TF position propensities and cell selectivity of conserved network motifs.** **a**, Shown is the propensity of all TFs within the ES cell regulatory network to occupy the different positions within a FFL. FFL positions are defined in panel **c**. **b**, Shown is the GO term enrichment of TFs that preferentially occupy position C within FFLs as opposed to TFs that preferentially occupy positions A and B within FFLs. Asterisk indicates a *q* value less than 0.05. *P* values and *q* values calculated using the Gene Ontology enrichment analysis and visualization tool (GOrilla). **c**, For all instances of FFLs in mouse ES cells, shown is the tissue specificity of each component edge across the other 24 mouse cell types. *P* values were calculated using a Wilcoxon rank sum test. **d**, Same as **c** but for regulating mutual motifs.




**Extended Data Figure 8 | Polarity of TF genes and regulatory networks during development.** **a**, Schematic illustrating the definition of and contrasting effector-facing and TF-facing TFs. **b**, Top: a box-and-whisker plot shows the distribution of the relative log enrichment of TF-facing to effector-facing TFs in mouse ES cells. Bottom: relative target landscape enrichments for individual TFs grouped together based on their functional categories. **c**, Shown is the GO term enrichment of TFs that preferentially regulate TFs (TF-facing) as opposed to TFs that preferentially regulate effector genes (effector-facing). Asterisk indicates a *q* value less than 0.05. *P* values and *q* values calculated using the Gene Ontology enrichment analysis and visualization tool (GOrilla). **d**, For each cell type, shown is the average propensity of each TF within the regulatory network to regulate TF genes versus effector genes. Relative enrichment values were calculated such that 0 indicates a cell-type regulatory network that is equally geared towards regulating TF genes and effector genes. Cell types are grouped/coloured according to their developmental origin. *P* values were calculated using a Wilcoxon rank sum test. **e**, Same as **b** but for human iPS cells. For box-and-whisker plots, box represents mean ± 25% and 75% quartiles, whiskers represent minimum and maximum values excluding outliers, and outliers indicated by open circles are defined as values outside 1.5 times the interquartile range above the upper quartile and bellow the lower quartile.

**Extended Data Table 1 | Baseline DNase I characteristics of the different mouse cell types**

| Cell Type | Stam ID | GEO Accession | GEO description | Sequenced reads | % of tags in DHSs | DNaseI Footprints |
|---|---|---|---|---|---|---|
| Activated Regulatory T-Cells | DS20149 | GSM1003834 | UW_DnaseDgf_TReg-Activated_adult-8wks | 349,952,959 | 56.57% | 874,813 |
| Activated TH0 T-Cells | DS17070 | GSM1003833 | UW_DnaseDgf_THelper-Activated_adult-8wks | 371,822,116 | 58.08% | 1,219,070 |
| Adipose Tissue (Genital) | DS18182 | GSM1014173 | UW_DnaseSeq_GenitalFatPad_adult-8wks_C57BL/6 | 429,875,952 | 56.73% | 2,810,616 |
| B-cell Lymphoma (A20) | DS16695 | GSM1003829 | UW_DnaseDgf_A20_immortalized | 295,681,721 | 50.76% | 871,180 |
| B-lymphocytes (blood) | DS16168 | GSM1003814 | UW_DnaseDgf_B-cell_(CD19+)_adult-8wks | 322,193,809 | 50.88% | 776,914 |
| B-lymphocytes (spleen) | DS17866 | GSM1003813 | UW_DnaseDgf_B-cell_(CD43-)_adult-8wks | 295,375,241 | 54.24% | 514,668 |
| Brain Tissue | DS12727 | GSM1003823 | UW_DnaseDgf_WholeBrain_adult-8wks | 224,580,229 | 70.93% | 1,019,584 |
| Erythroleukemia (MEL) | DS13036 | GSM1003824 | UW_DnaseDgf_MEL_immortalized | 314,608,167 | 58.18% | 1,083,560 |
| ES Cells (mCJ7) | DS13320 | GSM1003830 | UW_DnaseDgf_ES-CJ7_E0 | 266,022,035 | 49.30% | 623,778 |
| ES Cells (ZhBTc4 Oct4 KO 24hr) | DS17562 | GSM1003821 | UW_DnaseDgf_ZhBTc4_E0_DS17562 | 308,580,836 | 53.79% | 806,057 |
| ES Cells (ZhBTc4 Oct4 KO 6hr) | DS15236 | GSM1014150 | UW_DnaseSeq_ZhBTc4_E0_diffProtB_6hr_129/Ola | 367,428,781 | 57.94% | 1,111,148 |
| ES cells (ZhBTc4) | DS17616 | GSM1003822 | UW_DnaseDgf_ZhBTc4_E0_DS17616 | 289,624,956 | 58.58% | 814,349 |
| Fetal Brain Tissue | DS14536 | GSM1003828 | UW_DnaseDgf_WholeBrain_E14.5 | 343,697,514 | 61.68% | 1,409,418 |
| Fibroblast (NIH-3T3) | DS16900 | GSM1003831 | UW_DnaseDgf_NIH-3T3_immortalized | 382,389,955 | 50.99% | 830,004 |
| Heart Tissue | DS18138 | GSM1003820 | UW_DnaseDgf_Heart_adult-8wks | 415,035,272 | 54.23% | 1,459,061 |
| Kidney Tissue | DS13948 | GSM1003819 | UW_DnaseDgf_Kidney_adult-8wks | 234,471,226 | 57.07% | 992,665 |
| Liver Tissue | DS14605 | GSM1003818 | UW_DnaseDgf_Liver_adult-8wks | 221,364,696 | 71.71% | 1,107,823 |
| Lung Tissue | DS14479 | GSM1003817 | UW_DnaseDgf_Lung_adult-8wks | 380,969,896 | 58.55% | 1,560,827 |
| Mammary Adenocarcinoma | DS8497 | GSM1003816 | UW_DnaseDgf_3134_immortalized | 190,035,895 | 70.11% | 703,657 |
| Myeloid Progenitors (CD34+) | DS14099 | GSM1003815 | UW_DnaseDgf_416B_immortalized | 272,786,298 | 60.76% | 991,022 |
| Resting Regulatory T-Cells | DS17864 | GSM1003826 | UW_DnaseDgf_TReg_adult-8wks | 390,387,826 | 63.49% | 673,251 |
| Resting TH0 T-Cells (indiv. 1) | DS16171 | GSM1003825 | UW_DnaseDgf_T-Naive_adult-8wks | 346,731,260 | 56.83% | 1,005,818 |
| Resting TH0 T-Cells (indiv. 2) | DS17080 | GSM1003825 | UW_DnaseDgf_T-Naive_adult-8wks | 397,225,296 | 56.95% | 1,018,060 |
| Retina Tissue | DS20004 | GSM1003832 | UW_DnaseDgf_Retina_newborn-1days | 355,990,288 | 53.78% | 763,124 |
| Thymus Tissue | DS18819 | GSM1003827 | UW_DnaseDgf_Thymus_adult-8wks | 300,315,031 | 50.77% | 738,854 |

Database and sequencing information for the 25 mouse cell types used in this study.

**Extended Data Table 2 | Orthologous mouse and human cell types used for in-depth analyses**

| | Mouse cell type | | | Orthologous human cell type | | |
|---|---|---|---|---|---|---|
| Cell Type | Stam ID | GEO Acc. | GEO description | Stam ID | GEO Acc. | GEO description |
| ES Cells | DS17616 | GSM1003822 | UW_DnaseDgf_ZhBTc4_E0_DS17616 | DS11909 | GSM510581 | X_Hs_hESCT0_E_091028_02_DS11909_W |
| Brain | DS12727 | GSM1003823 | UW_DnaseDgf_WholeBrain_adult-8wks | DS11872 | GSM723021 | UW.Fetal_Brain.Digital_Genomic_Footprinting |
| Fibroblast | DS16900 | GSM1003831 | UW_DnaseDgf_NIH-3T3_immortalized | DS11923 | Accession | X_Hs_NHDFneo_E_091028_04_DS11923_W |
| Hematopoietic Progenitor | DS14099 | GSM1003815 | UW_DnaseDgf_416B_immortalized | DS12274 | GSM723022 | UW.Mobilized_CD34_Primary_Cells.Digital_Ge... |
| Erythroleukemia | DS13036 | GSM1003824 | UW_DnaseDgf_MEL_immortalized | DS16924 | ENCODE3-pending | |
| Thymus | DS18819 | GSM1003827 | UW_DnaseDgf_Thymus_adult-8wks | DS20341 | GSM1027351 | UW.Fetal_Thymus.Digital_Genomic_Footprinting |
| Treg Cells | DS17864 | GSM1003826 | UW_DnaseDgf_TReg_adult-8wks | DS14702 | GSM1014523 | UW_DnaseDgf_Treg_Wb78495824 |
| B-lymphocyte | DS16168 | GSM1003814 | UW_DnaseDgf_B-cell_(CD19+)_adult-8wks | DS18208 | GSM1014525 | UW_DnaseDgf_CD20+_RO01778 |

Description and database information for the orthologous mouse and human cell types used for various analyses in this study.