Supplemental Information

EXTENDED EXPERIMENTAL PROCEDURES

Regulatory Network Construction

We mapped motif-binding protein information found in TRANSFAC to 538 coding genes, using GeneCards (Rebhan et al., 1997) and UniProt Knowledgebase (Magrane and Consortium, 2011). Due to database annotations, some of these 538 coding genes were indistinguishable, as multiple genes were annotated as binders to the same set of motif templates by TRANSFAC. In such cases, we chose a single gene, randomly, as a representative and removed others. This reduced the number of genes from 538 to 475. Additionally, we included in this final set motif models for SOX2, OCT4, and KLF4 from the JASPAR Core database (Bryne et al., 2008).

We symmetrically padded the TSSs of these 475 genes by 5 kb and scanned for predicted TRANSFAC motif-binding sites using FIMO (Bailey et al., 2009), version 4.6.1, with a maximum p value threshold of 1×10^{-5} and defaults for other parameters. For each cell type, we filtered putative motif binding sites to those that overlapped footprints by at least 3 nt using BEDOPS (Neph et al., 2012b) as previously described (Neph et al., 2012a). Each network contained 475 nodes, one per gene. A directed edge was drawn from a gene node to another when a motif instance, potentially bound by the first gene's protein product, was found within a DNaseI footprint contained within 5 kb of the second gene's TSS, indicating regulatory potential. Table S3 shows the number of edges in every cell-type-specific network.

An approximately 150 nt region of duplicated sequence in the proximal regulatory region of the *NANOG* gene, with high sequence similarity to a single region proximal to a nearby *NANOG* pseudogene, prevented many DNaseI-seq reads from mapping per our usual procedure. To identify DNaseI footprints within this central promoter site, we mapped all non-uniquely-mappable reads falling within ± 5 kb of the TSS of the *NANOG* gene in each cell type. We then performed standard footprint detection on this region as previously described (Neph et al., 2012a), except that we did not filter footprints with >20% of its length covering non-uniquely-mappable locations. TF-binding elements within these DNaseI footprints were included in our final networks.

Network Visualization

We identified interactions that were unique to a single cell type, or "cell specific," and marked those found in two or more of the 41 tested cell types as "common." Interactions were rendered with Circos (Krzywinski et al., 2009), version 0.55. Within Circos nomenclature, two pseudo-chromosomes (ideograms) represent identically sorted lists of "regulator" and "regulated" factors, with a directed edge between ideograms indicating that the first factor regulates the second. Ideograms were colored by association of the cell type with tissue category. Unique and common interactions between ideograms were labeled with yellow and black colors, respectively, to visually differentiate cell types by the number and distribution of edges. TFs were oriented along both ideograms by the sort order provided by the H7-hESC cell type, from highest degree (SP1) to lowest (ZNF354C) (Table S1). For the detail view of H7-hESC, we also highlighted the interactions of four pluripotent (KLF4, NANOG, POU5F1, SOX2) and four constitutive factors (SP1, CTCF, NFYA, MAX) with purple and green edges, respectively.

Hive Plots

We generated a hive plot (Krzywinski et al., 2011) using the R library HiveR, version 0.2.1, to visualize directed interactions for four hematopoietic (PU.1, TAL1, ELF1, GATA2) and four pluripotent factors (KLF4, NANOG, OCT4, SOX2) among six cell types (H7-hESC, HRCEpiC, CD34+, HMVEC_dBlNeo, fBrain, and HSMM). The hive plot was divided into six sections, one for each cell type. Reading the figure in clockwise fashion, a directed edge drawn from one axis to the next indicates the first gene regulating the second. Genes were oriented identically along each axis. Common interactions were defined by an interaction existing in two or more cell types. A second qualitative hive plot was created between the same six cell types and over all 475 TFs (Table S1).

Unique Edge Connectedness

We calculated the mean weakly connected component size using edges unique to a cell type (Figures S1D–S1F and Table S2). To identify whether these unique component subnetworks were more connected than would be expected by chance, we randomly subsampled the same number of real edges in the same cell type and recalculated the mean-component size. This process was iterated 100,000 times, and the number of times for a cell type that the mean-component size in random graphs equaled or exceeded that of the unique component graph counterpart was tallied. An empirical p value was calculated as the tally plus one divided by 100,000. Subnetworks made up of unique edges belonging to each of HSMM, HRCEpiC, and H7-hESC were separately plotted using Cytoscape (Figures S1D–S1F) (Smoot et al., 2011).

Network Clustering

We counted the total number edges for every TF gene node (sum of in and out edges) in a cell type and calculated the proportion of edges for that TF relative to all edges in that cell type (NND). We computed the pairwise euclidean distances between cell types using the rescaled NND vectors and grouped the cell types using Ward clustering (Ward, 1963). We observed similar cluster patterns when comparing rescaled in-degree, rescaled out-degree, or unscaled total degree (results not shown).

Cell-Type-Specific Behaviors

We utilized the mfinder software (Milo et al., 2004), version 1.20, to pull out all FFL instances in regulatory networks. Prior to using the software, all self-edges, those from a TF gene node to itself, were removed per the requirements of the software. The software parameters were set to *-ospmem < motif-number > -maxmem 1000000 -s 3 -r 250 -z −2000*, where *< motif-number >* was one of 13 possible unique three-node network motif identifiers.

Triad Significance Profiles

We removed self-edges from every network and used the mfinder software tool for network motif analysis (Milo et al., 2004). 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 to estimate a null. We vectorized z-scores from every cell type and normalized each to unit length to create TSP as described in Milo et al. (2004). We computed the average TSP over all cell-type-specific regulatory networks and compared to the TSP of the highly curated multicellular information processing networks described in Milo et al. (2004). All sum squared error (SSE) calculations were done by comparing our derived networks against the *Caenorhabditis elegans* profile (White et al., 1986) (Table S3).

To generate a transcriptional network using only motif scan predictions we created a new network, with 86,242 edges, by using all putative motifs within 5 kb of the TSSs of each of the 475 TF genes, without conditioning on footprint overlaps. We analyzed this network using the mfinder software as described above, creating a TSP and comparing to the *Caenorhabditis elegans* profile.

To generate a transcriptional network from DNaseI footprints from all cell types we merged footprints across all cell types and filtered motif instances to those overlapping the merged set by at least 3 nt using BEDOPS (Neph et al., 2012b), creating another new network with 38,165 edges. We analyzed this network using the mfinder software as described above, creating a TSP and comparing to the *Caenorhabditis elegans* profile.

Network Feature Overlaps

We compared cell-type-specific networks in greater detail using only FFLs. Summaries of overlaps were made between a small number of cell types using Venn diagrams and barplots. All pairwise overlaps were computed and summarized using the Jaccard index (number of FFLs in the pairwise set intersection divided by the number in the pairwise set union—Figure S3E). We additionally computed overlaps and differences between entire regulatory networks in terms of shared and unshared edges, as well as footprints (Figures S1B and S1C).

To identify the contribution of each factor to each network motif, we counted the number of times a factor was present in each of the 13 three-node network motifs within the H7-hESC cell type, in any motif position (Figure S3F). We scaled each column vector to length 100, and then divided each element of a row vector by the maximum value in that row to visualize contributions in heatmap form using the matrix2png program without row normalization (Pavlidis and Noble, 2003).

SUPPLEMENTAL REFERENCES

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. *37* (*Web Server issue*), W202–W208.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res. *19*, 1639–1645.

Krzywinski, M., Birol, I., Jones, S.J., and Marra, M.A. (2011). Hive plots–rational approach to visualizing networks. Briefings in Bioinformatics. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22155641 [Accessed December 12, 2011].

Magrane, M., and Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. Database: The Journal of Biological Databases and Curation *2011*, bar009.

Neph, S.J., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman, R.E., Sandstrom, R., Johnson, A.K., Humbert, R., et al. (2012a). An expansive human regulatory lexicon encoded in transcription factor footprints. Nature. http://dx.doi.org/10.1038/nature11212.

Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Reynes, E., Maurano, M.T., Vierstra, J., Thomas, S., Sandstrom, R. Humbert, R., and Stamatoyannopoulos, J.A. (2012b). BEDOPS: high-performance genomic feature operations. Bioinformatics *28*, 1919–1920.

Pavlidis, P., and Noble, W.S. (2003). Matrix2png: a utility for visualizing matrix data. Bioinformatics *19*, 295–296.

Rebhan, M., Chalifa-Caspi, V., Prilusky, J., and Lancet, D. (1997). GeneCards: integrating information about genes, proteins and diseases. Trends Genet. *13*, 163.

Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.-L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics *27*, 431–432.

Ward, J.H. (1963). Hierarchical Grouping to Optimize an Objective Function. J. Am. Stat. Assoc. *58*, 236.

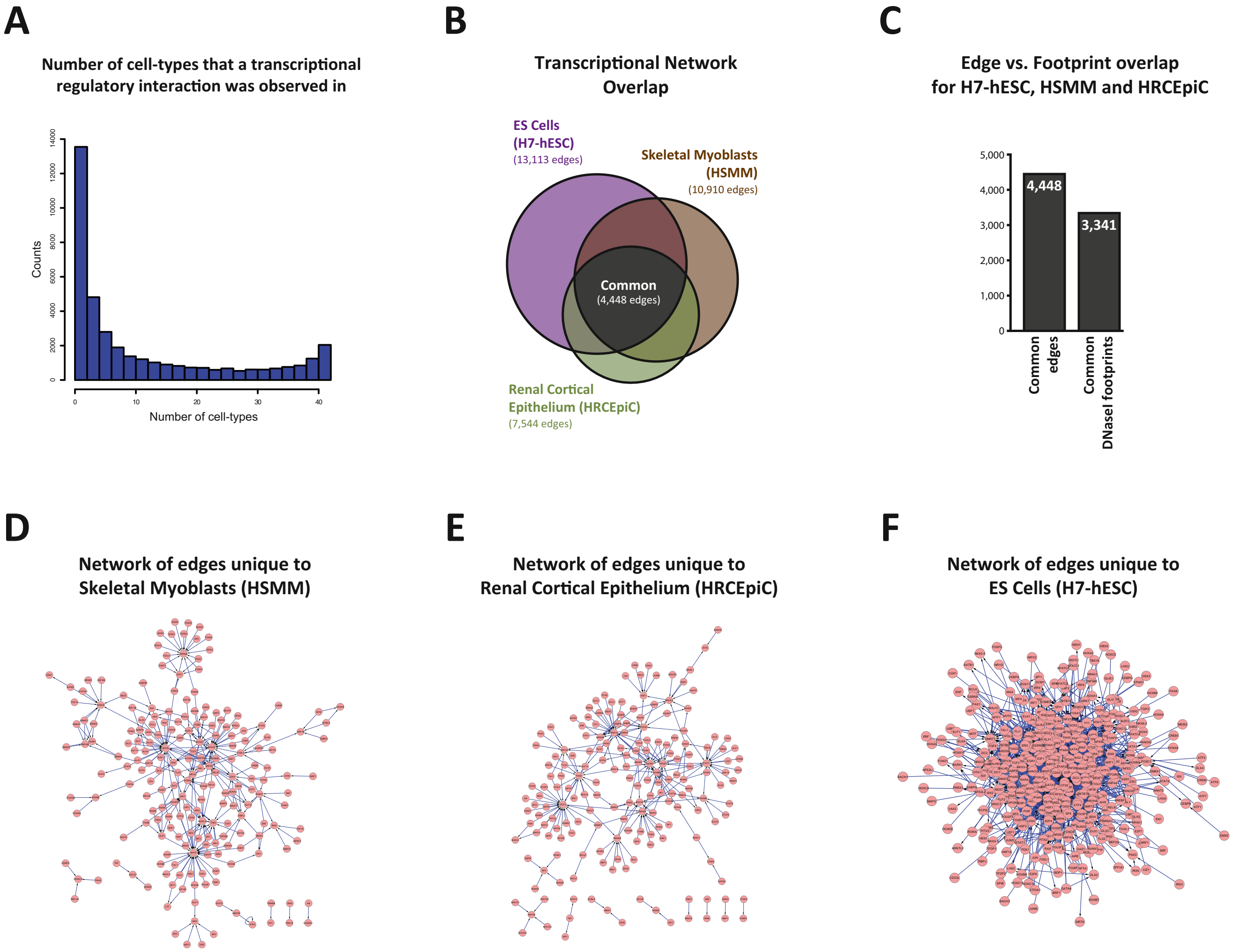

Figure S1. Overlap of Cell-Type-Specific Transcriptional Regulatory Networks, Related to Figure 3

(A) Histogram showing the number of cell types that each transcriptional regulatory interaction (edge) was observed in.

(B) The overlap of transcriptional regulatory interactions (edges) identified in ESCs (H7-hESC), skeletal muscle myoblasts (HSMM), and renal cortical epithelium (HRCEpiC).

(C) The number of common edges and common DNaseI footprints between the ESCs (H7-hESC), HSMM, HRCEpiC networks.

(D) Cytoscape derived network showing all edges that are unique to the HSMM network.

(E) Cytoscape-derived network showing all edges that are unique to the HRCEpiC network.

(F) Cytoscape-derived network showing all edges that are unique to the ESC (H7-hESC) network.

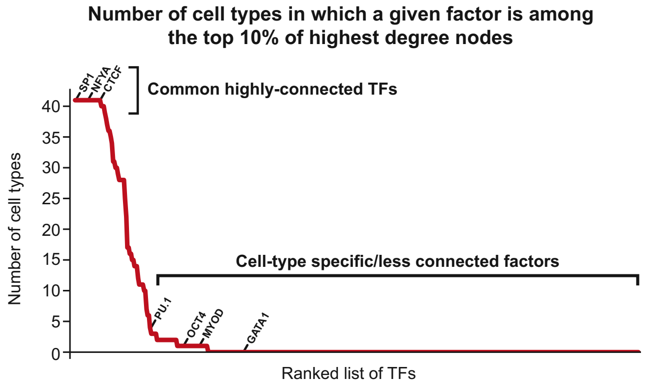

Figure S2. Identification of Common Highly Connected TFs, Related to Figure 4

Shown is the number of cell-type-specific networks in which a given factor is among the top 10% of highest degree nodes.

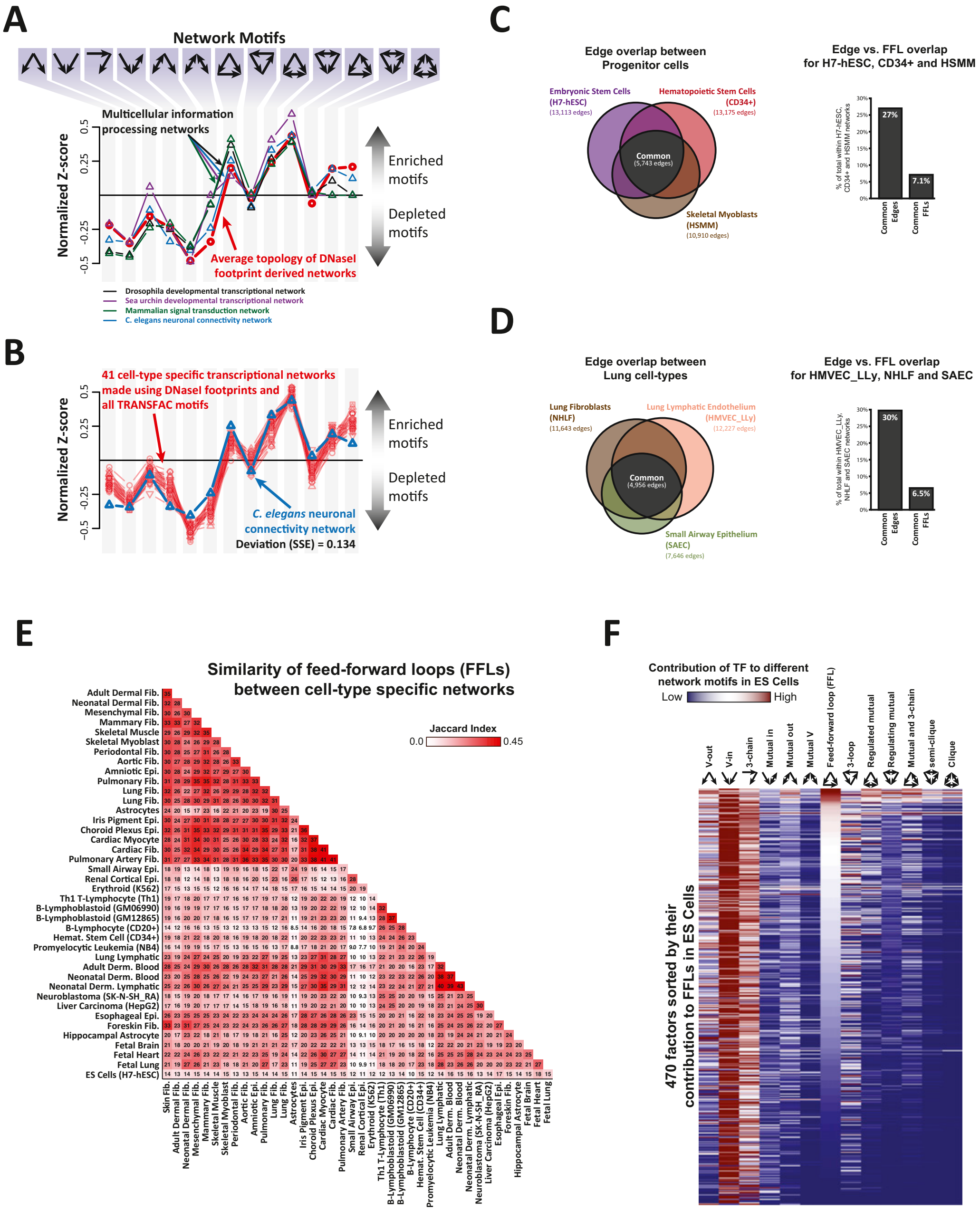

Figure S3. Transcriptional Regulatory Networks Have a Conserved Network Motif Architecture, Related to Figure 6

(A) Shown is the average relative enrichment or depletion of the 13 possible three-node architectural network motifs within the regulatory networks of each cell type (red line), compared with the relative enrichment of the same motifs in four previously published multicellular biological networks (Milo et al., 2004); *C. elegans* neuronal connectivity network (blue line), the mammalian signal transduction network (green line), and the sea-urchin (purple line) and *Drosophila* (black line) developmental transcriptional networks.

(B) Shown is the relative enrichment or depletion of the 13 possible three-node architectural network motifs within the regulatory networks of each cell type constructed using all 538 TRANSFAC motifs, including redundant motifs (red lines).

(C) The overlap of edges identified in three progenitor cell types—ESCs (H7-hESC), hematopoietic stem cells (CD34$^+$), and HSMM. Shown to the right is the percentage of all edges common to these three cell types, as well as the percentage of all FFLs common to these three cell types.

(D) The overlap of edges identified in three pulmonary cell types—NHLFs, HMVEC_LLy cells, and SAECs. Shown to the right is the percentage of all edges common to these three cell types, as well as the percentage of all FFLs common to these three cell types.

(E) Overlap of FFLs from networks of each cell type, following the ordering shown in Figure 4A. The color of each box corresponded to the Jaccard index between FFLs from the two cell-type-specific networks contributing to that box.

(F) Heatmap showing the contribution of all 470 TFs with interactions in ESCs (H7-hESC) to 13 possible three-node architectural network motifs in the ESC-type-specific network. The factors are sorted by their contribution to FFLs.