

An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man

Timothy Ravasi,^{1,4,5,22} Harukazu Suzuki,^{1,2,3,6,22} Carlo Vittorio Cannistraci,^{1,4,5,7,8,9,22} Shintaro Katayama,^{1,2,6,22} Vladimir B. Bajic,^{1,5,10,22} Kai Tan,^{1,4,23} Altuna Akalin,^{1,11} Sebastian Schmeier,^{1,10} Mutsumi Kanamori-Katayama,^{1,2,6} Nicolas Bertin,^{1,2,6} Piero Carninci,^{1,2,6} Carsten O. Daub,^{1,2,6} Alistair R.R. Forrest,^{1,2,6,12} Julian Gough,^{1,13} Sean Grimmond,^{1,14} Jung-Hoon Han,^{1,15} Takehiro Hashimoto,^{1,2,6} Winston Hide,^{1,10,16} Oliver Hofmann,^{1,10} Hideya Kawaji,^{1,2,6} Atsutaka Kubosaki,^{1,2,6} Timo Lassmann,^{1,2,6} Erik van Nimwegen,^{1,17} Chihiro Ogawa,^{1,2,6} Rohan D. Teasdale,^{1,14} Jesper Tegnér,^{1,18,19} Boris Lenhard,^{1,11} Sarah A. Teichmann,^{1,15} Takahiro Arakawa,^{1,2,6} Noriko Ninomiya,^{1,2,6} Kayoko Murakami,^{1,2,6} Michihira Tagami,^{1,2,6} Shiro Fukuda,^{1,2,6} Kengo Imamura,^{1,2,6} Chikatoshi Kai,^{1,2,6} Ryoko Ishihara,^{1,2,6} Yayoi Kitazume,^{1,2,6} Jun Kawai,^{1,2,6} David A. Hume,^{1,20} Trey Ideker,^{1,4,21,*} and Yoshihide Hayashizaki^{1,2,3,6,*}

¹The FANTOM Consortium

²RIKEN Omics Science Center

³General Organizers

⁴Departments of Medicine and Bioengineering

University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

⁵Red Sea Integrative Systems Biology Laboratory, Division of Chemical & Life Sciences and Engineering, Computational Bioscience Research Center, King Abdullah University for Science and Technology, Jeddah, Kingdom of Saudi Arabia

⁶RIKEN Omics Science Center, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho Tsurumi-ku Yokohama, Kanagawa, 230-0045 Japan

⁷Department of Mechanics, Politecnico di Torino, I-10129 Turin, Italy

⁸Proteome Biochemistry, San Raffaele Scientific Institute, 20132 Milan, Italy

⁹CMP Group Microsoft Research, Politecnico di Torino, I-10129 Turin, Italy

¹⁰South African National Bioinformatics Institute, University of the Western Cape, Private Bag X17, Bellville, 7535 South Africa

¹¹Bergen Center for Computational Science, Høyteknologisenteret Thormøhlensgate 55, N-5008 Bergen, Norway

¹²The Eskitis Institute for Cell and Molecular Therapies, Griffith University, QLD 4111, Australia

¹³Department of Computer Science, University of Bristol, Merchant Venturers Building, Woodland Road, Bristol, BS8 1UB, UK

¹⁴Australian Research Council Special Research Centre for Functional and Applied Genomics, Institute for Molecular Bioscience, The University of Queensland, St. Lucia, QLD 4072, Australia

¹⁵MRC Laboratory of Molecular Biology, Cambridge CB2 0QH, UK

¹⁶Biostatistics Department, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, USA

¹⁷Biozentrum, University of Basel, and Swiss Institute of Bioinformatics, Klingelbergstrasse 50/70, CH-4056 Basel, 4056, Switzerland

¹⁸Computational Medicine Group, Atherosclerosis Research Unit, Center for Molecular Medicine, Department of Medicine, Karolinska Institutet, Karolinska University Hospital Solna SE- 171 76 Stockholm, Sweden

¹⁹Department of Physics, Chemistry and Biology, Linköping University, SE-581 83 Linköping, Sweden

²⁰The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Roslin, EH259PS, UK

²¹The Institute for Genomic Medicine, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

²²These authors contributed equally to this work

²³Present address: Department of Internal Medicine & Biomedical Engineering, 285 Newton Road, University of Iowa, Iowa City, IA 52242, USA

*Correspondence: tideker@ucsd.edu (T.I.), yoshihide@gsc.riken.jp (Y.H.)

DOI 10.1016/j.cell.2010.01.044

SUMMARY

Combinatorial interactions among transcription factors are critical to directing tissue-specific gene expression. To build a global atlas of these combinations, we have screened for physical interactions among the majority of human and mouse DNA-binding transcription factors (TFs). The complete networks contain 762 human and 877 mouse interactions. Analysis of the networks reveals that highly connected TFs are broadly expressed across tissues, and that roughly half of the measured interactions are conserved between mouse and human. The data

highlight the importance of TF combinations for determining cell fate, and they lead to the identification of a SMAD3/FLI1 complex expressed during development of immunity. The availability of large TF combinatorial networks in both human and mouse will provide many opportunities to study gene regulation, tissue differentiation, and mammalian evolution.

INTRODUCTION

Tissue specificity is enabled by spatial and temporal patterns of gene expression which in turn are driven by transcriptional regulatory networks (Naef and Huelsken, 2005; Zhang et al., 2004).

Such networks involve assemblies of control proteins, such as DNA-binding transcription factors (TFs) connected to the sets of promoters of genes they induce or repress (Tan et al., 2008b). Typically, TFs do not act independently but form complexes with other TFs, chromatin modifiers, and cofactor proteins, which bind together and assemble upon the regulatory regions of DNA to affect transcription (Fedorova and Zink, 2008). Mapping the combinatorial interactions among TFs would represent a significant leap forward in our understanding of how tissue specificity is determined.

In recent years, a variety of genome-scale technologies have been introduced which allow mammalian transcriptional regulatory networks to be investigated at high resolution and depth. Many such studies have inferred transcriptional networks through mRNA expression profiling combined with genome-wide active promoter mapping and promoter motif analysis (e.g., Suzuki et al., 2009). These data have been supplemented with fluorescence-activated cell sorting (FACS) (Shachaf et al., 2008) or reverse transcriptase quantitative polymerase chain reaction (qRT-PCR) (Roach et al., 2007; Wen et al., 1998).

Another technology that has revolutionized the study of transcriptional networks is chromatin immunoprecipitation (ChIP), which when coupled with microarrays or high-throughput sequencing (Johnson et al., 2007), enables genome-wide measurements of TF binding locations in vivo. A complementary approach is the protein binding microarray (PBM) (Berger et al., 2008), which rapidly characterizes the complete DNA sequence repertoire bound by a TF in vitro. ChIP and PBMs have been applied to map transcriptional networks in a variety of human cell types, including stem cells (Cole et al., 2008; Lee et al., 2006) and lymphocytes (Marson et al., 2007; Schreiber et al., 2006), and to characterize the binding motifs of many mammalian TF families (Berger et al., 2008).

Although these studies have led to the construction of very large models of transcriptional networks, they are based on experiments that largely treat each TF in isolation. For instance, ChIP-chip measures binding locations for one TF at a time, although separate profiles for several TFs can be later combined into networks (Mathur et al., 2008). However, it is well known that the transcriptional output of a gene is due to the joint activity of many TFs whose binding and activation are highly interdependent. This cooperation is often mediated by direct physical contact between two or more TFs, forming homodimers, heterodimers, or larger transcriptional complexes. In fact, it has been estimated that approximately 75% of all metazoan TFs heterodimerize with other factors (Walhout, 2006). Newman and Keating used protein arrays to reveal a network of several hundred domain interactions among the bZIP TF family alone (Grigoryan et al., 2009). Other studies have successfully assembled large networks of protein interactions using technologies such as coimmunoprecipitation and two-hybrid screening (Park et al., 2005; Yu et al., 2008), but to date these have not been systematically applied to map networks of transcription factors. Thus, a clear and immediate task is to map which combinations of TFs act together and how these combinations lead to modes of regulation that are not evident when each factor is considered separately.

Toward this goal, we have pursued an integrative approach to systematically map combinatorial interactions among mammalian

TFs. Our approach draws from two systems-wide data sets generated in both human and mouse: physical protein-protein interaction among TFs measured using the mammalian two-hybrid (M2H) system and quantitative TF expression levels measured across tissues by qRT-PCR. Analysis of these data identifies a database of TF complexes and networks that can be used to elucidate the regulatory programs behind developmental processes and disease. Chief among these results is a network of homeobox TFs, which we show can predict tissue type in mammals.

RESULTS

Mammalian Transcription Factor Protein-Protein Interaction Networks

We compiled a list of 1988 human and 1727 mouse DNA-binding transcription factors using information from public gene databases (Table S1). Of these, 1222 and 1112 cDNA clones were captured, in human and mouse, respectively, that could be verified to express full-length protein (Table S1). All pair-wise combinations of TF cDNAs were systematically screened for protein-protein interaction using the M2H system (Suzuki et al., 2001). Bait and prey constructs were cotransfected in CHO-K1 cells, and the interaction of the expressed proteins was monitored by luciferase reporter activity. This process identified 762 and 877 high-stringency TF-TF interactions in human and mouse, respectively (Tables S2 and S3). The use of M2H meant that the human and mouse TF interactions were measured in near-physiological conditions including mammalian posttranslational and other modifications. The web-accessible atlas of all pairwise TF interactions mapped by M2H is available at <http://fantom.gsc.riken.jp/4/tf-ppi>. This resource is searchable by gene ID or function and provides network visualizations as well as raw lists of interactions.

To estimate the sensitivity of the screening approach (the percentage of all true TF-TF interactions that are identifiable by M2H), we assembled a gold-standard set of high-confidence TF-TF dimers reported in previous literature. To obtain this gold standard, a set of 289 mouse TF-TF interactions were downloaded from public databases and further curated to select 91 interactions supported by two or more independent lines of evidence or primary experimental reports (Supplemental Information and Table S3). We found that M2H recovered protein-protein interactions for 23 of these heterodimers, yielding a sensitivity of 25%. Apart from sensitivity, we were also interested in precision (the percentage of reported interactions that are true, equal to 1 – false discovery rate). Precision is more difficult to estimate than sensitivity, because it requires a gold standard that contains not only known interactions but also a large number of protein pairs that are known to be noninteracting. Since such data are not available, we sought to confirm the M2H positives using in vitro pull-down assays as a second technology. Of 34 randomly chosen mouse M2H positives, 18 (53%) were detected by in vitro pull-down (Table S4). This second assay is not a gold standard, such that failure to confirm an M2H positive by in vitro pull-down does not negate the corresponding protein-protein interaction, which might be transient or unstable under conditions of the pull-down. However, this analysis does show that the M2H network recovers approximately one quarter of known TF heterodimers and that the majority of M2H interactions can be replicated by a second

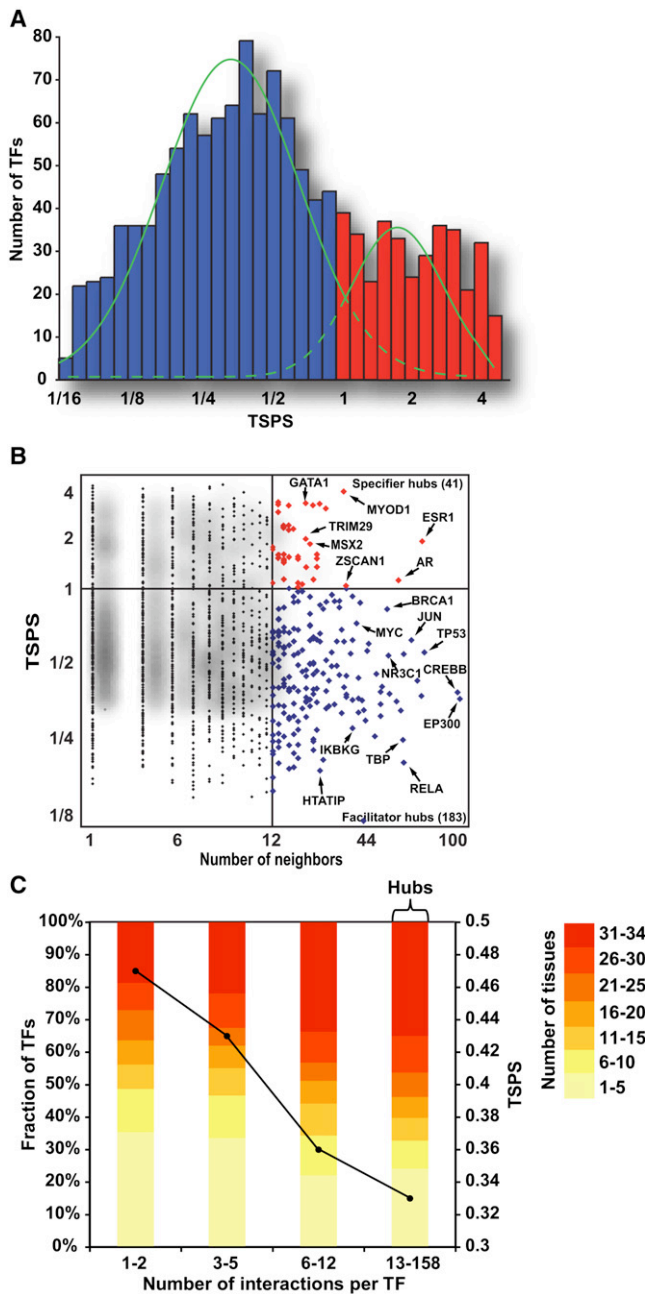


Figure 1. TF Expression versus Connectivity

(A) Distribution of tissue specificity for all TFs. The green curves fit the bi-modal distribution as a mixture of two Gaussians.

(B) Scatterplot of tissue specificity (y axis) versus number of neighbors (x axis). Red points are defined as specifier hubs and blue points as facilitator hubs (Table S1).

(C) TFs are binned into four groups of approximately equal size based on their number of interactions (x axis). The tissue specificity distribution of each bin is represented by stacks of colored segments. Segment height represents the fraction of TFs in an expression group (left y axis), and segment color represents the number of tissues in which TFs in that group are expressed. The black line displays the median TSPS of each group (right y axis). The results shown are for human M2H interactions supplemented with human TF-TF interactions downloaded from literature (Table S2); similar results are obtained for mouse

technology. These figures are consistent with high quality interaction networks published elsewhere recently (Yu et al., 2008).

We now describe four case studies that use the atlas to address questions of how transcriptional control contributes to tissue specificity in mammals. These case studies cover: (1) integration of the atlas with quantitative TF abundance levels across human and mouse tissues, revealing a prominent relationship between TF connectivity and expression; (2) identification of a subnetwork of homeobox factors that is highly discriminative and predictive of tissue type; (3) a proteome-wide map of conserved transcriptional complexes in mammals, many of which have tissue-specific expression patterns that are also highly conserved; and (4) examples of how the atlas can be used to recognize and further explore TF heterodimers in control of tissue differentiation.

Integration of TF Interaction and Expression Reveals Insights into Network Structure

In order to physically interact, TFs must be coexpressed in the same tissue or cell type. To investigate the tissue specificity of TF interactions, we obtained quantitative mRNA profiles of all TFs using qRT-PCR across a panel of 34 human and 20 mouse tissues (Table S5). For each TF, we computed a tissue-specificity score (TSPS), which uses relative entropy to quantify the extent to which the observed TF expression pattern departs from the null distribution of uniform expression across all tissues (Experimental Procedures, Table S1, and Table S5). Examination of tissue specificity over all TFs suggested a mixture of two distinct TF populations, with one population of TFs having widespread tissue expression (TSPS < 1) and a second smaller population at higher tissue specificity (TSPS ≥ 1, Figures 1A and 1B). We called the TFs with widespread expression “facilitators,” based on the hypothesis that they facilitate transcriptional programs across many different tissues, and we called those with high-specificity tissue “specifiers.” For example, the TFs JUN and FOS, which form the AP-1 heterodimer, were classified as strong facilitators owing to low TSPS (average around 0.6; Table S5). This score is consistent with the classical view of AP-1 as a broad activator of expression in major cellular processes including differentiation, proliferation, and apoptosis (Ameyar et al., 2003). In contrast, many TFs with known roles in tissue differentiation were classified as “specifiers,” such as MYOD1, which regulates muscle development and members of the Paired box (Pax) TF family involved in tissue morphogenesis. The observed bimodal distribution of TF expression is in agreement with recent findings from a meta-analysis of publicly-available expression profiles in humans (Vaquerizas et al., 2009).

Examining the relationship between expression and interaction, we observed a strongly negative Pearson correlation of -0.79 between a TF’s number of protein interactions and its TSPS. That is, we found that TFs with few interactions tend to be expressed in a tissue-specific pattern while TFs with many interactions—so called network “hubs” (Jin et al., 2007;

interactions or for M2H interactions only (Table S3; see also Table S4 for confirmation of the M2H positives using in vitro pull down assays as a second technology).

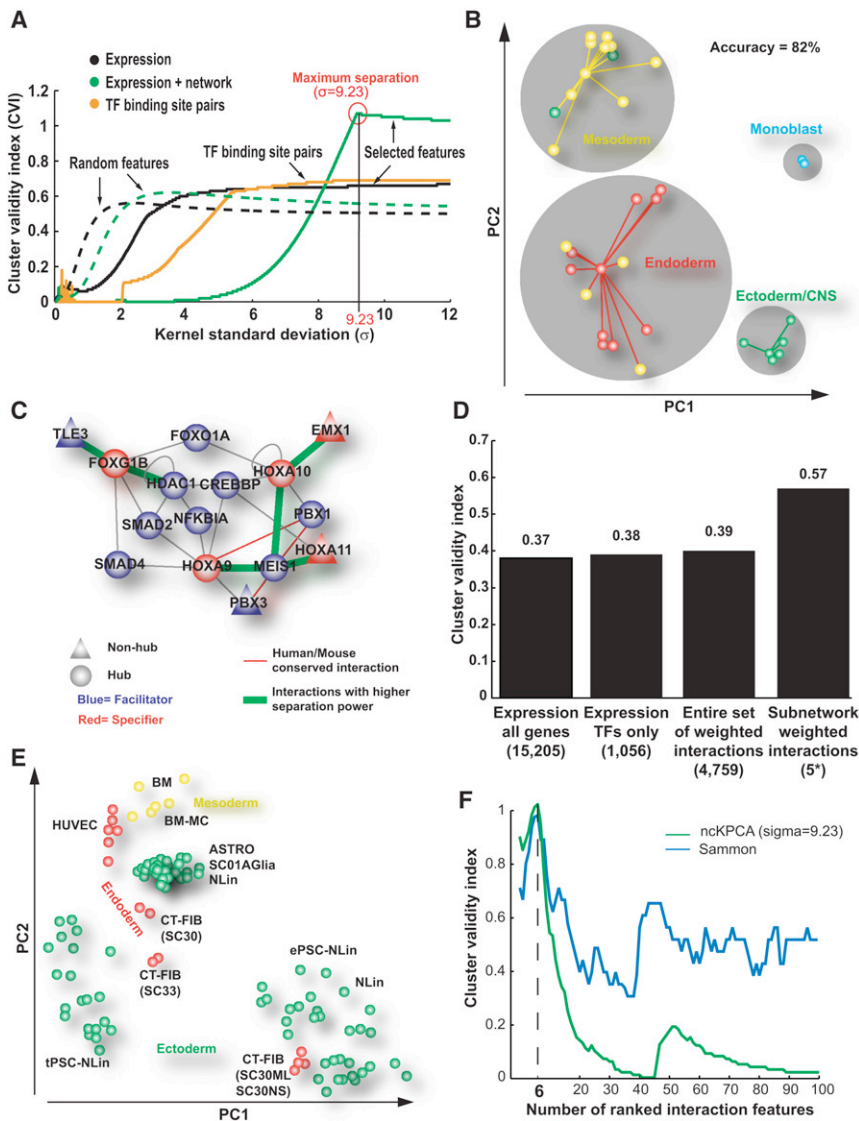


Figure 2. A Homeobox Network Associated with Tissue Differentiation

(A) Performance of tissue separation with (green solid curve) or without (black solid curve) information about TF protein-protein interactions (Table S2). The Bezdek cluster validity index (CVI, y axis) is a measure of separation between the four tissue classes. CVI is plotted for increasing kernel standard deviation (x axis), the only tuning parameter of the ncKPCA algorithm used for tissue separation. Performance was also evaluated for TF pairs predicted to cooperate based on co-occurrence of TF binding sites (yellow curve) (Yu et al., 2006) as well as for random features (dashed curves).

(B) Tissue dimensionality reduction by ncKPCA into the first two Principle Components (PCs), considering features derived from the six most informative TF-TF interactions. Points represent tissues derived from ectoderm (green), mesoderm (yellow), or endoderm (red), or a monocyte cell line (blue). Gray circles denote four clusters obtained by affinity propagation in the (PC1, PC2) space, with each point connected to its cluster exemplar. This figure is related to Figure S1.

(C) Informative subnetwork containing six interactions (green) used to generate features for tissue separation. Also shown are the immediate network neighbors of the interacting TFs.

(D) CVI for the separation of stem cells (Table S6) using Sammon Mapping. Four feature sets are shown: the original expression values from Muller et al., the expression of the TFs only, the entire set of TF protein-protein interactions, or the features corresponding to the six interactions in panel C (5* indicates that the interaction HOXA9-MEIS1 was not considered because HOXA9 expression was not measured in the stem cell investigation of Muller et al.).

(E) Stem cell dimensionality reduction obtained by Sammon Mapping using the panel C interaction set. Points represent stem cell lines derived from ectoderm (green), mesoderm (yellow), or endoderm (red).

(F) Good performance of tissue separation observed with two different algorithms. ncKPCA

(green curve) and Sammon Mapping (blue curve). CVI (y axis) is plotted against the number of PC2-ranked interactions used to separate tissues (x axis). In both cases, the maximum performance is observed using the first six PC2-ranked interactions to separate tissues.

Yu et al., 2006)—tend to be expressed across many tissues (Figure 1C). The observed correlation was highly significant, as assessed by 10,000 random trials in which the assignment of expression values to TFs was permuted ($r = 0.00 \pm 0.03$). Such widespread expression of TF hubs bears some similarity to previous studies of TF-DNA (transcriptional) interactions, in which the number of promoters bound by a TF was found to correlate with the number of growth conditions in which it is expressed (Luscombe et al., 2004; Zhou et al., 2008).

A Homeobox Network Associated with Specification of Tissue Type

Combinatorial interaction among transcription factors is critical for differentiation of tissues (Davidson et al., 2002). To identify TF interaction networks involved in tissue development, we clus-

tered the TF expression profiles across the 34 human tissues (see above) using two approaches: a basic tissue separation approach using expression levels only, and a “network-transformed” approach in which we exploited as features the differences in expression level across TF-TF interactions, as suggested by a recent study (Taylor et al., 2009). We found that network transformation resulted in an increased separation of tissues into four well-formed clusters (a 38% increase, Figures 2A and 2B and Figure S1). These corresponded to well-defined tissue classes according to embryonic origin: ectoderm (including central nervous system or CNS), mesoderm, endoderm, and cell lines. Strikingly, only six TF interactions were sufficient to classify tissue type with a high accuracy of 82% (Figures 2B and 2C). Moreover, we found that these interactions fell into the same small network neighborhood defined by a subnetwork

of 15 proteins (Figure 2C). This subnetwork was highly enriched for homeobox factors (7/15 proteins) many of which have, at least individually, known roles in tissue-type specification during development (Duverger and Morasso, 2008). Although we expected that many of these TFs would be tissue specifiers, we found that 10 of the 15 were in fact facilitators expressed broadly across most tissue types. These results support the notion that it is the interactions among transcription factors more than their expression levels alone that help to determine tissue identity.

Given the ability of the homeobox-related subnetwork to separate tissues based on their embryological origin, we sought to test whether this subnetwork was also able to discriminate the embryological origin of different types of stem cells. Understanding the transcriptional events that commit stem cells to different tissue lineages is one of the major goals of stem cell research (Jaenisch, 2009). For this purpose, we downloaded the publicly-available gene expression profiles of 219 stem cell lines derived from a variety of different tissue types (Muller et al., 2008) (Table S6 lists the tissue origin of each cell line). As shown in Figures 2D and 2E, the homeobox-related subnetwork was indeed able to separate these stem cell expression profiles by ectoderm, mesoderm, and endoderm origin. This separation was 33% better than that achieved using other methods (Figure 2D). This analysis suggests that the good performance of the homeobox-related subnetwork (Figure 2C) is not the result of overfitting to a specific set of tissue expression profiles. Moreover, it provides further evidence that the combinatorial interactions revealed in this subnetwork play an important role in cell commitment to different tissue lineages.

Conservation of TF Complexes across Mammalian Evolution

A strong line of evidence that a particular TF interaction is functional is observation of cross-species conservation of that interaction. For each human TF, we used the InParanoid algorithm (O'Brien et al., 2005) to identify its set of amino acid sequence orthologs in mouse. We then identified pairs of TFs for which the orthologs were observed to interact in both species. In total, 80 conserved interactions were identified between the M2H data of human and mouse—this number rose to 305 conserved interactions when supplementing M2H data with literature (Table S2 and Table S3). Considering this number together with the M2H sensitivity and precision estimates above, we computed the fraction of conserved TF-TF interactions between human and mouse to be in the range of 34%–64% (depending on the value one uses for the precision of M2H screening, see Supplemental Information).

We next used NetworkBLAST (Kalaev et al., 2008) to examine how these conserved interactions clustered within the network, i.e., whether they fell within common subnetworks suggestive of conserved transcriptional complexes. In total, 68 conserved complexes were identified which contained approximately six TFs on average. Examples of conserved complexes are shown in Figures 3A–3F; the complete set is included as part of the atlas at <http://fantom.gsc.riken.jp/4/tf-ppi>. Eighty percent of the conserved complexes were enriched for gene ontology biological process annotations. These conserved TF complexes

provide a first-draft map of the combinatorial regulatory circuits common to mammals.

The conserved complexes also suggest combinations of heterodimers in specific biological contexts for future investigation. Figure 3C shows a conserved complex of six TFs in which five are broadly expressed across all tissues in both species, and one TF (LHX2) is restricted to frontal cortex also in both species (Table S5). Figures 3D–3F show three conserved TF complexes consisting of proteins coexpressed in cerebellum. Messenger RNA in situ hybridization analysis of mouse cerebellum, obtained from the Allen Brain Atlas (Lein et al., 2007), confirms that the interacting TFs are indeed expressed in cerebellum and that this localization is cerebellum-specific at single-cell resolution.

FLI1 and SMAD3 Form a Heterodimeric Complex Associated with Monocyte Development

The vast majority of TF-TF interactions recorded in the atlas represent new combinations not yet documented in the literature. Thus, an important question is how particular interactions of interest should be carried forward in the laboratory to identify new transcriptional heterodimers and to study their regulatory functions. As an example use of the atlas to identify tissue-restricted heterodimers, four interactions were selected for which at least one TF had moderate to high tissue specificity (Figure 4A). For example, Peroxisome Proliferator-Activated Receptor Gamma (PPARG) is expressed in adipose, skin, lung, and breast, with little or no expression in other tissues. Although its interaction partner, Retinoid X Receptor Beta (RXRB), is expressed ubiquitously the interaction requires the presence of both TFs and thus remains tissue restricted (Table S5).

Given these tissue-restricted TF combinations, a first step was to characterize and further establish their physical interaction. We used bidirectional in vitro pull-down assays to examine whether each TF pair could exhibit strong, stable, and direct physical binding under the conditions of the pull-down, independent of other proteins or factors. As shown in Figure 4B, all four TF interactions were recapitulated as in vitro pull-downs, making them strong candidates for functional transcriptional complexes.

Next, we sought detailed information on the dynamic expression of a TF combination in the tissue(s) in which both TFs were active. One of the identified TF interactions was between Friend Leukemia virus Integration 1 (FLI1) and SMAD family member 3 (SMAD3), in which FLI1 was restricted primarily to macrophage-related tissues (THP-1, spleen, lymph) while SMAD3 was found to be expressed more generally (Figure 4A and Table S5). Thus, we investigated the role of the FLI1/SMAD3 interaction in macrophage differentiation, using qRT-PCR to record a time-course of expression of both TFs during differentiation of THP-1 monoblasts to monocytes following stimulation by PMA. Strikingly, both TFs were coordinately downregulated at early time points during differentiation (Figure 4C). These data are supported by previous findings in which SMAD3 has been shown to regulate cell proliferation through TGF- β 1 signaling (Meran et al., 2008), and FLI1 has been shown to reactivate NOTCH pathways resulting in p53-dependent cell-cycle arrest (Ban et al., 2008). A hypothesis for future work is that FLI1/SMAD3 may function together as a repressor complex that controls cell proliferation during differentiation (Figure 4D).

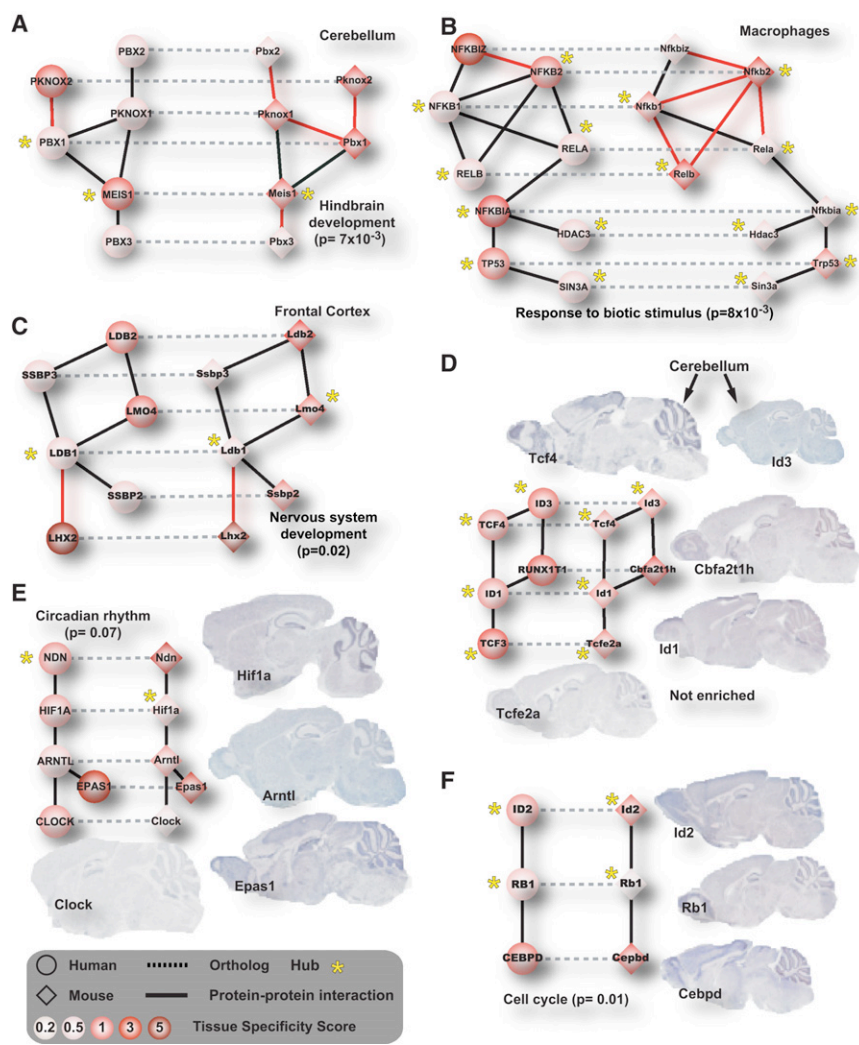


Figure 3. TF Subnetworks Conserved across Human and Mouse

(A–F) Examples of TF subnetworks conserved in specific tissues. Human proteins are circles and mouse proteins are diamonds, colored in increasing shades of red representing increasing tissue specificity (TSPS), (Table S1). Stars indicate hubs. Horizontal dashed links indicate protein orthology relationships across species, whereas solid links indicate protein-protein interactions within species (red links are newly-discovered, black links are literature-curated).

(D–F) Conserved TF subnetworks that are specific to cerebellum, as first indicated by qRT-PCR (red nodes and Table S5) and subsequently confirmed by in situ hybridization to mouse brain tissue samples. All conserved subnetworks are available at <http://fantom.gsc.riken.jp/4/tp-pi>.

(specifiers) tend to interact with TFs that are broadly-expressed (Figure 1), increasing the number of possible combinatorial events only in certain tissues or during tightly-regulated developmental processes. In support of this interaction-centric model, we identified a subnetwork of just 15 TFs that was sufficient to confer maximal separation of tissues and stem cell lines into the three germ layers associated with embryogenesis (Figure 2). This network significantly outperformed tissue separation based on the expression of individual factors alone. Two thirds of these “germ layer” factors were facilitator TFs expressed in the majority of tissues.

The theme of specificity through interaction is also evident among the conserved TF subnetworks (Figure 3).

The majority of TFs in these networks are broadly expressed, and it is the minority of TFs that confer tissue specificity. Further evidence comes from the four identified TF complexes we validated and placed into biological contexts (Figure 4 and Table S5). Although they were not selected on this basis, at least three of these complexes involve combination of a tissue restricted TF (i.e., NR3C1, PPARG, FLI1) with a partner whose expression pattern is more widespread (RXRB, RXRB, SMAD3).

The availability of large TF-TF combinatorial interaction networks in both human and mouse will provide many opportunities to study network conservation and divergence over the course of mammalian evolution. Debate is still ongoing regarding the rate at which various types of molecular networks evolve. Here, we found that conservation between human and mouse TF-TF interactions was moderate (Figure 3), in the range of 34 to 64 percent. In contrast, a recent comparison of transcriptional (protein-DNA) interactions reported that this type of network is highly divergent over even very short evolutionary timescales (Tuch et al., 2008). A comparison of genetic networks (synthetic lethal and epistatic interactions) also found extreme rates of

DISCUSSION

In this study, we have mapped an atlas of combinatorial interactions among the majority of human and mouse TFs. This work makes available a number of significant resources for the biomedical community, including a database of over 1600 human or mouse TF-TF interactions (Tables S2 and S3) and quantitative TF expression measurements across human and mouse tissues (Table S5). The data highlight conserved TF subnetworks whose patterns of interaction and tissue specificity suggest transcriptional complexes in control of tissue identity.

Our analysis, derived by the integration of these datasets, supports a model whereby the transcriptional network structure is dominated by facilitator TFs expressed broadly across tissues (Figure 1 and Table S1). The implication is that tissue identity is not determined by tissue-restricted TFs, but relies on tissue-restricted interaction among TFs. Each TF may be expressed in a variety of tissues, but it is only where two TFs are co-expressed and colocalized that an interaction, and its functional consequences, may occur. In this model, tissues restricted TFs

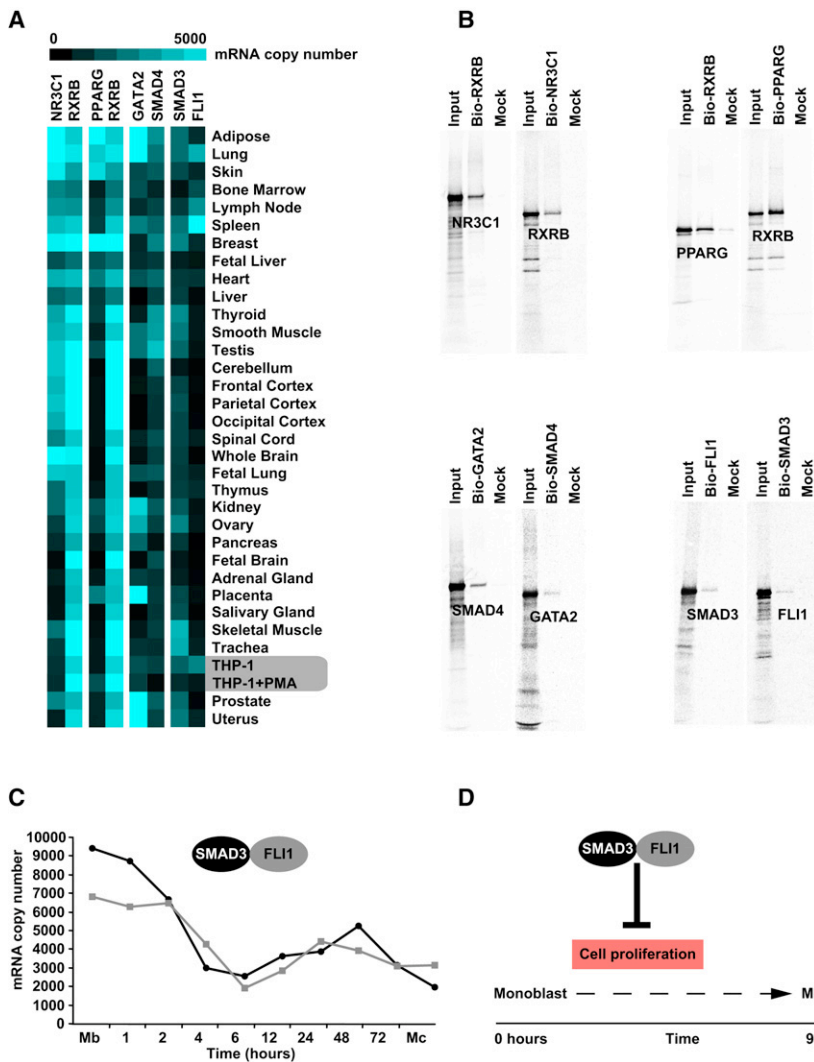


Figure 4. Physical and Functional Exploration of Tissue-Restricted Heterodimers

(A) Four heterodimers that display combinatorial logic across tissues. The heatmap shows the mRNA copy number of each heterodimeric TF across tissues measured by qRT-PCR (Table S5). (B) In vitro pull down experiment shows clear bidirectional physical interaction for each of the four heterodimers as detected originally by M2H assay (Table S2).

(C) mRNA levels of FLI1 and SMAD3 during THP-1 differentiation induced by PMA, as measured by qRT-PCR.

(D) Graphical representation of FLI1/SMAD3 control during myeloid differentiation.

EXPERIMENTAL PROCEDURES

Mammalian Two-Hybrid Assays

Following PCR amplification of full-length TFs, M2H was carried out as previously described (Usui et al., 2005). To assess potential for self-activation each BIND TF fragment (bait) was transfected into CHO-K1 cells containing the luciferase reporter plasmid pG5luc. Reporter activity was measured after 20 hr and BIND samples with high self-activation (more than 5x larger than average) were removed. For non-self-activating baits, eight BIND TF fragments (baits) and two ACT TF fragments (preys) were cotransfected into CHO-K1 cells with pG5luc, and luciferase reporter activity was measured after 20 hr. The screen was also performed using two BIND TFs combined with two ACT TFs. For transfections with positive reporter activity, the assay was repeated using all 2 x 2 or 8 x 2 BIND/ACT combinations to identify the interacting TF pairs. Positive interactions were scored as those that showed at least three times higher luciferase activity than background (measured using transfection of

either an ACT-TF or BIND-TF alone). For more details see Supplemental Information, Table S2, and Table S3.

In Vitro Pull-Down Assay

PCR products encoding the TF coding sequence and the SV40LPAS fragment were used to construct a template for in vitro transcription/translation. The products were combined by overlapping PCR using the primer pair T7-RBS-KOZAK (5'-GAGCGCGCGTAATACGACTCACTATAGGGGAAGGAGCCGCC ACCATG-3') and LGT10L (5'-AGCAAGTTTCAGCTGGTTAAG-3'), yielding a final template encoding a 5' T7 RNA polymerase promoter. In vitro pull-down assays were carried out as previously described (Suzuki et al., 2004). Briefly, biotinylated or [³⁵S]-labeled TF was synthesized in vitro from the template using Transcend Biotinylated lysine-tRNA (Promega) or Redivue L-[³⁵S]-methionine (Amersham Biosciences) in combination with the TNT T7 Quick Coupled Transcription/Translation System (Promega). After confirmation of [³⁵S]-labeled protein synthesis by SDS-PAGE and autoradiography, biotinylated protein and [³⁵S]-labeled protein were mixed 1:1 and incubated on ice for one hour. Control reactions containing [³⁵S]-labeled protein alone were conducted in parallel. The reaction was then incubated with streptavidin Dynabeads (DynaL Biotech, Milwaukee, WI) for 30 min at 4°C on a rotary shaker. Dynabeads were isolated with a magnet and washed 5 times with ice-cold TBST buffer (50 mM Tris-HCl [pH 8.0], 137 mM NaCl, 2.68 mM KCl, 0.1% Tween 20). The amount of radio-labeled protein coprecipitated with the biotinylated

divergence (Roguev et al., 2008). On the other hand, protein-protein interactions, especially those that form major structural and functional components of the eukaryotic cell, were found to be highly conserved (Tan et al., 2008a). Protein-protein interactions forming transcriptional complexes, as we have studied here, appear to be conserved at an intermediate level somewhere between the extremes. That is, TF-TF complexes are likely more mutable than the major complexes of cell structure and central metabolism, but much less so than the rapid rewiring that appears to take place in networks of transcription factor / promoter binding.

It has long been appreciated that gene regulation involves combinatorial interactions among transcription factors. The contribution of the present work is to map, on a global scale, precisely what many of these connections are. With few exceptions, almost all of the uncovered connections are undocumented in the existing literature. Future work will dissect more precisely how each of these combinations contributes to developmental programs and to an individual's relative state of health or disease.

protein was measured by scintillation counting or was detected by SDS-PAGE. The ratio of scintillations with and without biotinylated protein was calculated to measure the interaction between the two proteins (Table S4).

Tissue Specificity Score

The value f_j^i , the fractional expression level of TF i in tissue j , was computed as the ratio of the TF expression level in tissue j (qRT-PCR) to its sum total expression level across all tissues. Tissue specificity TSPS _{i} was then computed using relative entropy:

$$TSPS_i = \sum_j f_j^i \log_2 f_j^i / (q^i)$$

where q^i is the fractional expression of TF _{i} under a null model assuming uniform expression across tissues. According to this definition, a minimal TSPS = 0 would be reported for TFs expressed uniformly across all tissues, while a maximal TSPS \cong 5 would be reported for TFs expressed only in a single tissue. The threshold chosen for classifying TFs as tissue “specifiers” (TSPS \geq 1) was based on the observed bimodal distribution of expression over all TFs and tissues (Figure 1A). This threshold is conservative, as it selects TFs with roughly a 20-fold expression difference or greater across tissues (Tables S1 and S5).

Unsupervised Tissue Separation

Two different feature sets were considered for tissue separation: (1) TF expression values and (2) TF-TF interaction values. For both feature sets the raw qRT-PCR expression values were normalized so that each tissue had the same average value over all TFs, then log transformed (Tables S1 and S5). Following (Taylor et al., 2009) interaction values were computed for each interaction between a hub and any other TF, with hubs taken as TFs with > 12 interactions (Figure 1C, Table S2, and Table S3). Separations were performed using a hybrid two-phase procedure. The first phase was noncentered Principal Components Analysis (ncPCA), in which the second principal component resulting from this analysis (PC2) was found to be the main direction informative for tissue separation (either feature set). The features were then ranked according to their absolute PC2 loadings and a second phase of dimensionality reduction was performed using the ranked features. For this second phase, noncentered Kernel PCA (ncKPCA) was used with two parameters: (1) the standard deviation of the Gaussian kernel and (2) the number of top-ranked features selected for separation. Performance of separation into the tissue classes was measured by the Bezdek cluster validity index (CVI) considering the first two dimensions (PC1, PC2). Further details are provided in the Supplemental Information.

We also examined the dependence of tissue specification on the particular network used. Although the M2H network reported here (Tables S2 and S3) is the first large-scale experimental screen for TF-TF interactions, previous studies have sought to predict relevant TF combinations based on co-occurrence of TF binding sites within gene promoters (Yu et al., 2006). However, we found that a network of TF pairs predicted using binding site co-occurrence did not perform as well as the network of physical TF interactions elucidated by M2H and previous literature (Figure 2A). We also found that the performance of network-based tissue specification was not dependent on the particular algorithm used for separation. Both ncKPCA and Sammon Mapping approaches yielded very similar performance with Cluster Validity Index (CVI) \cong 1, and in both cases CVI was maximized for exactly six interactions (Figure 2F).

Data and Analysis Results

The data and analysis results of the paper are available from <http://fantom.gsc.riken.jp/4/tf-ppi>.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, one figure, and six tables and can be found with this article online at doi:10.1016/j.cell.2010.01.044.

ACKNOWLEDGMENTS

The work for the RIKEN Omics Science Center was supported by grants from the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) through the Genome Network Project and for the RIKEN Omics Science Center (YH, Principal Investigator). Members of the FANTOM Consortium were supported by grant MH062261 from the US National Institute of Mental Health (TR, KT, TI), the King Abdullah University of Science and Technology (TR, VBB), the Max Planck Society for the Advancement of Science (AK), the SA National Bioinformatics Network (SS, AR, VBB, WAH), the Claude Leon Foundation (MK), a CJ Martin Fellowship from the Australian NHMRC (ARRF), and the Scuola Interpolitecnica di Dottorato (CVC). The authors gratefully acknowledge S. Choi for critical feedback on the manuscript.

Competing interests' statement: The authors declare that they have no competing financial interests.

Received: June 6, 2009

Revised: September 22, 2009

Accepted: January 25, 2010

Published: March 4, 2010

REFERENCES

- Amezar, M., Wisniewska, M., and Weitzman, J.B. (2003). A role for AP-1 in apoptosis: the case for and against. *Biochimie* 85, 747–752.
- Ban, J., Bennani-Baiti, I.M., Kauer, M., Schaefer, K.L., Poremba, C., Jug, G., Schwentner, R., Smrzka, O., Muehlbacher, K., Aryee, D.N., et al. (2008). EWS-FLI1 suppresses NOTCH-activated p53 in Ewing's sarcoma. *Cancer Res.* 68, 7100–7109.
- Berger, M.F., Badis, G., Gehrke, A.R., Talukder, S., Philippakis, A.A., Pena-Castillo, L., Alleyne, T.M., Mnaimneh, S., Botvinnik, O.B., Chan, E.T., et al. (2008). Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* 133, 1266–1276.
- Cole, M.F., Johnstone, S.E., Newman, J.J., Kagey, M.H., and Young, R.A. (2008). Tcf3 is an integral component of the core regulatory circuitry of embryonic stem cells. *Genes Dev.* 22, 746–755.
- Davidson, E.H., Rast, J.P., Oliveri, P., Ransick, A., Caestani, C., Yuh, C.H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., et al. (2002). A genomic regulatory network for development. *Science* 295, 1669–1678.
- Duverger, O., and Morasso, M.I. (2008). Role of homeobox genes in the patterning, specification, and differentiation of ectodermal appendages in mammals. *J. Cell. Physiol.* 216, 337–346.
- Fedorova, E., and Zink, D. (2008). Nuclear architecture and gene regulation. *Biochim. Biophys. Acta* 1783, 2174–2184.
- Grigoryan, G., Reinke, A.W., and Keating, A.E. (2009). Design of protein-interaction specificity gives selective bZIP-binding peptides. *Nature* 458, 859–864.
- Jaenisch, R. (2009). Stem cells, pluripotency and nuclear reprogramming. *J. Thromb. Haemost.* 7 (Suppl 1), 21–23.
- Jin, G., Zhang, S., Zhang, X.S., and Chen, L. (2007). Hubs with network motifs organize modularity dynamically in the protein-protein interaction network of yeast. *PLoS ONE* 2, e1207.
- Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316, 1497–1502.
- Kalaev, M., Smoot, M., Ideker, T., and Sharan, R. (2008). NetworkBLAST: comparative analysis of protein networks. *Bioinformatics* 24, 594–596.
- Lee, T.I., Jenner, R.G., Boyer, L.A., Guenther, M.G., Levine, S.S., Kumar, R.M., Chevalier, B., Johnstone, S.E., Cole, M.F., Isono, K., et al. (2006). Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* 125, 301–313.
- Lein, E.S., Hawrylycz, M.J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A.F., Boguski, M.S., Brockway, K.S., Byrnes, E.J., et al. (2007). Genome-wide atlas of gene expression in the adult mouse brain. *Nature* 445, 168–176.

- Luscombe, N.M., Babu, M.M., Yu, H., Snyder, M., Teichmann, S.A., and Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* *431*, 308–312.
- Marson, A., Kretschmer, K., Frampton, G.M., Jacobsen, E.S., Polansky, J.K., MacIsaac, K.D., Levine, S.S., Fraenkel, E., von Boehmer, H., and Young, R.A. (2007). Foxp3 occupancy and regulation of key target genes during T-cell stimulation. *Nature* *445*, 931–935.
- Mathur, D., Danford, T.W., Boyer, L.A., Young, R.A., Gifford, D.K., and Jaenisch, R. (2008). Analysis of the mouse embryonic stem cell regulatory networks obtained by ChIP-chip and ChIP-PET. *Genome Biol.* *9*, R126.
- Meran, S., Thomas, D.W., Stephens, P., Enoch, S., Martin, J., Steadman, R., and Phillips, A.O. (2008). Hyaluronan facilitates transforming growth factor-beta1-mediated fibroblast proliferation. *J. Biol. Chem.* *283*, 6530–6545.
- Muller, F.J., Laurent, L.C., Kostka, D., Ulitsky, I., Williams, R., Lu, C., Park, I.H., Rao, M.S., Shamir, R., Schwartz, P.H., et al. (2008). Regulatory networks define phenotypic classes of human stem cell lines. *Nature* *455*, 401–405.
- Naef, F., and Huelsenken, J. (2005). Cell-type-specific transcriptomics in chimeric models using transcriptome-based masks. *Nucleic Acids Res.* *33*, e111.
- O'Brien, K.P., Remm, M., and Sonnhammer, E.L. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* *33*, D476–D480.
- Park, D., Lee, S., Bolser, D., Schroeder, M., Lappe, M., Oh, D., and Bhak, J. (2005). Comparative interactomics analysis of protein family interaction networks using PSIMAP (protein structural interactome map). *Bioinformatics* *21*, 3234–3240.
- Roach, J.C., Smith, K.D., Strobe, K.L., Nissen, S.M., Haudenschild, C.D., Zhou, D., Vasicek, T.J., Held, G.A., Stolovitzky, G.A., Hood, L.E., et al. (2007). Transcription factor expression in lipopolysaccharide-activated peripheral-blood-derived mononuclear cells. *Proc. Natl. Acad. Sci. USA* *104*, 16245–16250.
- Roguev, A., Bandyopadhyay, S., Zofall, M., Zhang, K., Fischer, T., Collins, S.R., Qu, H., Shales, M., Park, H.O., Hayles, J., et al. (2008). Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science* *322*, 405–410.
- Schreiber, J., Jenner, R.G., Murray, H.L., Gerber, G.K., Gifford, D.K., and Young, R.A. (2006). Coordinated binding of NF-kappaB family members in the response of human cells to lipopolysaccharide. *Proc. Natl. Acad. Sci. USA* *103*, 5899–5904.
- Shachaf, C.M., Gentles, A.J., Elchuri, S., Sahoo, D., Soen, Y., Sharpe, O., Perez, O.D., Chang, M., Mitchel, D., Robinson, W.H., et al. (2008). Genomic and proteomic analysis reveals a threshold level of MYC required for tumor maintenance. *Cancer Res.* *68*, 5132–5142.
- Suzuki, H., Forrest, A.R., van Nimwegen, E., Daub, C.O., Balwiercz, P.J., Irvine, K.M., Lassmann, T., Ravasi, T., Hasegawa, Y., de Hoon, M.J., et al. (2009). The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat. Genet.* *41*, 553–562.
- Suzuki, H., Fukunishi, Y., Kagawa, I., Saito, R., Oda, H., Endo, T., Kondo, S., Bono, H., Okazaki, Y., and Hayashizaki, Y. (2001). Protein-protein interaction panel using mouse full-length cDNAs. *Genome Res.* *11*, 1758–1765.
- Suzuki, H., Ogawa, C., Usui, K., and Hayashizaki, Y. (2004). In vitro pull-down assay without expression constructs. *Biotechniques* *37*, 918–920.
- Tan, K., Feizi, H., Luo, C., Fan, S.H., Ravasi, T., and Ideker, T.G. (2008a). A systems approach to delineate functions of paralogous transcription factors: role of the Yap family in the DNA damage response. *Proc. Natl. Acad. Sci. USA* *105*, 2934–2939.
- Tan, K., Tegner, J., and Ravasi, T. (2008b). Integrated approaches to uncovering transcription regulatory networks in mammalian cells. *Genomics* *91*, 219–231.
- Taylor, I.W., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, C., Faria, D., Bull, S., Pawson, T., Morris, Q., and Wrana, J.L. (2009). Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.* *27*, 199–204.
- Tuch, B.B., Li, H., and Johnson, A.D. (2008). Evolution of eukaryotic transcription circuits. *Science* *319*, 1797–1799.
- Usui, K., Katayama, S., Kanamori-Katayama, M., Ogawa, C., Kai, C., Okada, M., Kawai, J., Arakawa, T., Carninci, P., Itoh, M., et al. (2005). Protein-protein interactions of the hyperthermophilic archaeon *Pyrococcus horikoshii* OT3. *Genome Biol.* *6*, R98.
- Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A., and Luscombe, N.M. (2009). A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* *10*, 252–263.
- Walhout, A.J. (2006). Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping. *Genome Res.* *16*, 1445–1454.
- Wen, X., Fuhrman, S., Michaels, G.S., Carr, D.B., Smith, S., Barker, J.L., and Somogyi, R. (1998). Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci. USA* *95*, 334–339.
- Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., et al. (2008). High-quality binary protein interaction map of the yeast interactome network. *Science* *322*, 104–110.
- Yu, X., Lin, J., Zack, D.J., and Qian, J. (2006). Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. *Nucleic Acids Res.* *34*, 4925–4936.
- Zhang, W., Morris, Q.D., Chang, R., Shai, O., Bakowski, M.A., Mitsakakis, N., Mohammad, N., Robinson, M.D., Zirngibl, R., Somogyi, E., et al. (2004). The functional landscape of mouse gene expression. *J. Biol.* *3*, 21.
- Zhou, L., Ma, X., and Sun, F. (2008). The effects of protein interactions, gene essentiality and regulatory regions on expression variation. *BMC Syst. Biol.* *2*, 54.

EXTENDED EXPERIMENTAL PROCEDURES

Mammalian Transcription Factor Gene Lists and cDNA Libraries

A list of human TFs was constructed by combining four sources of information: the set of genes included in the TRANSFAC transcription factor database (Matys et al., 2006), the set of genes annotated as “transcription factor” by the Gene Ontology (GO) database (2008), the set of genes containing the word “transcription” in the Entrez Description field, and a manually-curated TF list (Roach et al., 2007). After combining these sources and removing redundancies, further manual curation was performed to remove proteins inappropriately included in the list, for a final count of 1988 proteins. A corresponding list of mouse TFs was constructed from the human list for a final count of 1727 proteins. The final human and mouse TF lists are provided in Table S1.

A collection of full length cDNAs for human TFs (1222 out of 1988 proteins) was obtained from the Mammalian Gene Collection (<http://mgc.nci.nih.gov/>), the FLJ cDNA Project (<http://fldb.hgc.jp>), and by additional RT-PCR cloning clones with verified sequences (988 MGC clones, 130 FLJ clones and 104 in house RT-PCR cloning clones). Mouse transcription factor full length cDNAs (1112 out of 1727 proteins) were obtained from the FANTOM cDNA library (Carninci et al., 2005). These cDNA collections were used for mammalian two-hybrid screens in human and mouse as detailed below (Table S1).

Mammalian Two-Hybrid Assays

In preparation for mammalian two-hybrid (M2H), gene-specific forward and reverse primers were used to PCR amplify each cDNA as follows. Each 38 bp forward primer was given an 18-bp consensus tag (5'-GAAGGAGCCGCCACCATG-3') followed by the first 20 bases of the protein coding sequence (CDS) in the sense orientation. Similarly, each 41 bp reverse primer was given a 21 bp consensus tag (5'-CAATTTACACAGGAACTCA-3') followed by the last 20 bases of the CDS in the anti-sense orientation. When the CDS length was >1500 bps, the CDS was assayed separately in equal length of fragments so that each amplified fragment has 300 bps of margin to the neighbor fragment and does not exceed 1650 bps.

Fragments for the CMV promoter and either the Gal4 DNA-binding domain (BIND) or the VP16 transcriptional activation domain (ACT) were PCR-amplified from pBIND or pACT vectors (Promega, Madison, Wisconsin) using the primers FPCMV6 (5'-CCAATATGACCGCCATGTTGGC-3') and RSALSE (5'-CATGGTGGCGGCTCCTTCAAGCAACGCGTCAAGTCGACT-3'). The fragment for the SV40 late polyadenylation signal (SV40LPAS) was PCR-amplified from the pBIND vector using the primer pair FSV40LPAS02 (5'-GTTTCTGTGTGAAATTGTTATCCGCTGCAGACATGATAAGATACATTG-3') and RSV40LPAS01 (5'-AGCAAGTTCAGCCTGGTTAATGATCCTTATCGATTTTACCAC-3'). Overlapping PCR was carried out to connect the CDS fragments with the BIND- or ACT-fragments and the SV40LPAS fragment using the primer pair FPCMV5 (5'-GCCATGTTGGCATTGATTATTGAC-3') and LGT10L (5'-AGCAAGTTCAGCCTGGTTAAG-3'). The sequences in bold underline are complementary to the common tag sequences of the CDS-specific forward and reverse primers, respectively. PCR conditions were based on those in our previous reports (Suzuki et al., 2004). All PCR products were verified by agarose gel electrophoresis after amplification.

Following PCR, M2H was carried out as previously described (Usui et al., 2005). Briefly, to assess potential for self-activation each BIND TF fragment (bait) was transfected into CHO-K1 cells containing the luciferase reporter plasmid pG5luc. Reporter activity was measured after 20 hr and BIND samples with high self-activation (with more than 5-fold larger than average) were removed. For non-self-activating baits, eight BIND TF fragments (baits) and two ACT TF fragments (preys) were cotransfected into CHO-K1 cells with pG5luc, and luciferase reporter activity was measured after 20 hr. The screen was also performed using two BIND TFs combined with two ACT TFs. For transfections with positive reporter activity, the assay was repeated using all 2 × 2 or 8 × 2 BIND/ACT combinations to identify the interacting TF pairs. Positive interactions were scored as those that showed at least three times higher luciferase activity than background (measured using transfection of either an ACT-TF or BIND-TF alone), Tables S2 and S3.

Literature-Curated Interactions and Interaction Annotation

We collected known protein-protein interactions between TFs from the open-access databases BIND, DIP, HPRD, IntAct and MINT, considering *only those interactions* derived from low-throughput assays. This resulted in 4566 (human) and 289 (mouse) nonredundant, published interactions. When available, PubMed identifiers of publications reporting the interactions as well as the different detection methods are listed (Tables S2 and S3). For all pairs of interacting TFs, the GO semantic similarity of the Biological Process and Cellular Component annotations was calculated using the R Bioconductor package GOSemSim with the Wang similarity measure (Wang et al., 2007) (Tables S2 and S3). Similarity values near or equal to one indicate that the interacting TFs have a similar or equal GO annotation. The 91 mouse interactions comprising the gold standard list used to assess the sensitivity of M2H screening have been manually curated and each publication manually checked for consistency and reliability.

Quantitative RT-PCR

Total RNA extracted from human tissues was purchased from CloneTech, Stratagene, and Ambion. Preparation of total RNA from THP-1 samples was described in Carninci et al. (Carninci et al., 2006). All total RNA samples were verified to be free of contaminating genomic DNA, using real-time PCR with and without reverse transcription using primer pairs for ubiquitously-expressed genes including glyceraldehyde-3-phosphate dehydrogenase (GAPDH), β 2-macroglobulin (β 2M), and phosphoglycerate kinase (PGK). Gene-specific primer pairs were designed using the Primer3 software (<http://frodo.wi.mit.edu/primer3/>) with an optimal primer size of 20 bases, amplification size of 140 bp, and annealing temperature of 60°C. Primer pairs were located within exons so that

transcript number could be calculated using genomic DNA standards. For the vast majority of TF genes (>99%), primer pairs were designed within the final exon, which contains 3' noncoding sequences. Primer pairs in other exons were used < 1% of the time. The reasons for choosing the final exon are as follows: (1) The final exon is typically the longest exon, yielding a longer sequence for design of good primer pairs. (2) The final exon sequence is typically low in GC content, and (3) Microarray probes, such as Illumina arrays, are also typically designed preferentially against the final exons. The amplification efficiency of each primer pair was checked and, if poor, the primer pair was designed again. In cases in which it was impossible to design any more probes in the final exon, we selected another (long) exon (<1% of genes). Primer pairs used to quantitate human transcription factors are available at <http://fantom.gsc.riken.jp/4/tf-ppi>. First-strand cDNA synthesis (5 µg total RNA per 20 µl reaction) and PCR amplification were performed in triplicate with an ABI Prism 7900HT instrument (Applied Biosystems) as previously reported (Suzuki et al., 2004). Dissociation curve analysis was performed in accordance with the manufacturer's protocol, to ensure the amplification of a single DNA product. Expression profiles were obtained as Ct values which were converted into copy numbers by using coefficients (slope and intercept) of standard curves obtained using genomic DNA at defined amounts (1, 10, and 100 ng). Raw copy numbers were filtered to remove values < 40 and then normalized as previously described (Mar et al., 2006) using a standard curve of 18S rRNA measurements (Table S5).

Calculation of Overlap between Human and Mouse M2H Networks

In the Results section of the main paper, we report a conservation rate of 34% to 64% between the human and mouse M2H interactions (Tables S2 and S3). This rate is computed given the following quantities with respect to the human M2H interaction data set:

Observed positive human interactions: 762

Observed positive human interactions, considering baits whose orthologs were screened as baits in mouse: $O^+ = 502$

Observed positive human interactions, subset conserved with mouse: $O^c = 80$

False Negative Rate from gold standard: $\beta = 0.75$ *

False Discovery Rate from co-IPs: $0 < q < 0.47$ **

*In the manuscript, we use a gold-standard set of interactions to estimate $\beta = 0.75$, equal to $1 - \text{true positive rate}$.

**We place an upper bound on $q < 0.47$, since 53% [i.e., $1 - 0.47$] of a representative sample of M2H interactions could be validated by co-IP, while failure to co-IP is not strong negative evidence against an M2H interaction.

Goal: Compute A^c / A^+ , the fraction of actual positive interactions (A^+) that are conserved (A^c). The Actual numbers are different from their Observed counterparts which are affected by false negative and false discovery rates.

Solution: It can be shown that the fraction of actual conserved interactions is related to the fraction of observed conserved interactions by the formula:

$$\frac{A^c}{A^+} = \frac{O^c(1 - q)}{O^+(1 - \beta)}$$

For low false discovery rate $q \rightarrow 0$, this gives $A^c / A^+ = 64\%$. For high false discovery rate $q \rightarrow 0.47$, this gives $A^c / A^+ = 34\%$.

The above formula assumes that the same β and q apply to all observations in both species and these error rates are also conditionally independent between the two species. These assumptions may be false, i.e., it is possible that the chance of a true positive in mouse is dependent on whether the interaction has already been observed in humans. However, such dependencies are nontrivial to estimate and it is not clear in which direction they would influence the estimated fraction of conservation. Also, note that in the comparative analysis we use a very stringent criteria to define the interactions conserved in both species, that is, we considered only those evolutionary relationships between two TFs with a clear 1 to 1 ortholog. Therefore, we lose some "true" conserved interactions in particular between those TFs that have several paralogs in one or both species.

Tissue Separation: Feature Normalization and Transformation

The original TF expression values were normalized and log-transformed:

$$e_{ij}^* = -\log \frac{e_{ij}}{\sum_j e_{ij}}$$

where i identifies the tissue and j the TF. The size of the data vectors (i.e., dimensionality of the data space) is as follows. For Feature Set 1 (expression values, Table S5), the size is equal to the total number of human transcription factor genes that were interrogated by qPCR and were connected in the protein-protein interaction network (Table S2), yielding 1321 features. For Feature Set 2 (interaction weights) we used the set of interactions obtained from the literature or M2H, resulting in a total of 5227 features, Table S2 (collection of differential e^* values across each TF interaction "incident to a hub") by the method suggested by Taylor et al. (Taylor et al., 2009). For Feature Set (3), we used a set of 215 interactions among TF pairs predicted to cooperate based on co-occurrence of TF binding sites according to the method of Yu et al. (Yu et al., 2006).

Tissue Separation: Further Algorithmic Details

We used a hybrid (two-phase) procedure for tissue separation. Hybrid procedures are common in the field of machine learning, especially for speech and image processing as well as other applications including computational biology (Simon, 2004). The first phase of separation was performed with noncentered Principal Components Analysis (ncPCA). The second principal component resulting from this analysis (PC2) was found to be the main direction informative for tissue separation. Strikingly, this PC2 was found to be the most informative for all feature sets examined (expression features or interaction features). The features were then ranked according to their absolute PC2 loadings (i.e., the relative weight contribution of each feature to PC2). Next, a second round of dimensionality reduction was performed using the ranked features, for two reasons. First, we found that tissue separation could be further improved by use of a nonlinear machine learning approach based on the PC2 ranked in the first (linear) phase. Second, we wished to de-noise the features and to investigate if a particular subset of features (ranked in the first phase based on their importance for separation) could offer interesting biological insights. For the second round of separation, noncentered Kernel PCA (ncKPCA) was used with two parameters. The first parameter was the standard deviation of the Gaussian kernel— this was scanned over a range from 0.01 to 50 in steps of 0.01. The second parameter was the number of top-ranked features selected for separation - this was increased in steps of 1, starting from the first two ranked features and expanding to include all features. Performance of separation into the tissue classes was measured by the Bezdek cluster validity index (CVI) considering the first two dimensions (PC1, PC2). The maximum separation was achieved considering the first six interactions obtained by Feature Set 2 (interaction features).

To assess significance of the selected features, we ran the same separation procedure on equally-sized sets of features selected randomly. This result is reported in Figure 2A (dashed lines) in the main paper. In addition, we performed two analyses to address robustness of the separation approach. First, the same separation procedure was performed using Sammon Mapping (Sammon, 1969), an alternative parameter-free machine. As we show in Figure 2F, Sammon mapping obtained very similar results to ncKPCA, the approach used initially. The importance of this result is that Sammon mapping *has no parameters*, suggesting that our separation achieved with either approach is a property intrinsic to the biological system and is not an artifact of some particular choice of parameters. Second, we found that the six interaction features used for separation of tissues also performs very well for separation of a completely different data set – stem cell lines derived from the article of Muller et al. (Muller et al., 2008) (Figures 2D and 2E and Table S4). This analysis suggests that the detected six interactions resulting in the homeobox-related subnetwork are not the result of overfitting to a specific set of tissue expression profiles.

SUPPLEMENTAL REFERENCES

- Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al. (2005). The transcriptional landscape of the mammalian genome. *Science* 309, 1559–1563.
- Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A., Taylor, M.S., Engstrom, P.G., Frith, M.C., et al. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* 38, 626–635.
- Mar, J.C., Rubio, R., and Quackenbush, J. (2006). Inferring steady state single-cell gene expression distributions from analysis of mesoscopic samples. *Genome Biol.* 7, R119.
- Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* 34, D108–D110.
- Sammon, J.W. (1969). A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.* 18, 401–409.
- Simon, P. (2004). *Bioinformatics: The Machine Learning Approach*, 2nd edn, (Cambridge: Cambridge University Press).
- Suzuki, H., Okunishi, R., Hashizume, W., Katayama, S., Ninomiya, N., Osato, N., Sato, K., Nakamura, M., Iida, J., Kanamori, M., et al. (2004). Identification of region-specific transcription factor genes in the adult mouse brain by medium-scale real-time RT-PCR. *FEBS Lett.* 573, 214–218.
- Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S., and Chen, C.F. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics* 23, 1274–1281.

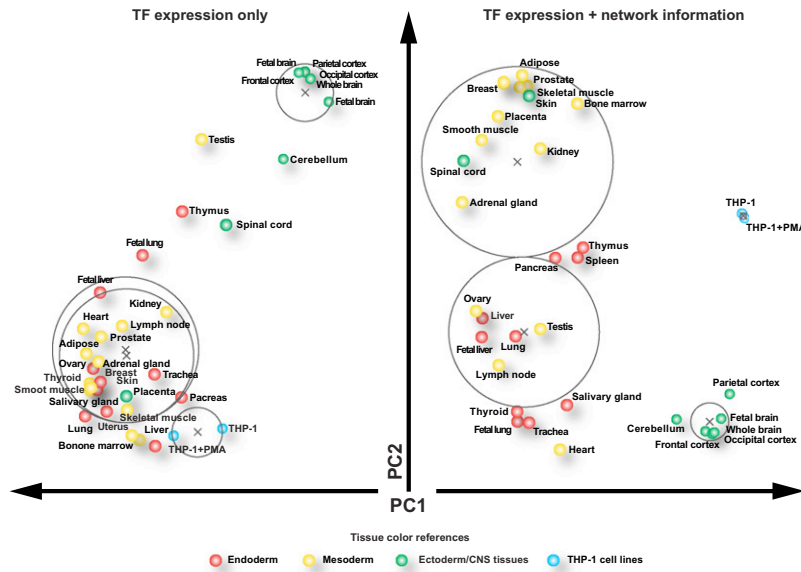


Figure S1. Tissue Cluster Separation in the 2D Reduction Space, Related to Figure 2B

Left, tissue dimensionality reduction for KPCA standard deviation = 9.23, considering TF expression only (Table S5). Right, tissue dimensionality reduction for KPCA standard deviation = 9.23, using the network transformed approach (Tables S2, S3, S4, and S5). The x axis is the first principal component (PC1), and the y axis is the second principal component (PC2). Grey crosses indicate cluster centroids. Grey circles indicate cluster scatter evaluated as average deviation from the centroid. Figure 2 shows the same tissues dimensionality reduction analysis with the distance of each tissue from the cluster centroid.