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
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).
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 ± 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 clustered 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 (Kalai 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-β1 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).
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 coexpressed and colocalized that an interaction, and its functional consequences, may occur. In this model, tissues restricted TFs (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...
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.

**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 pGL3uc. Reporter activity was measured after 20 hr and BIND samples with high self-activation (more than 5× 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 pGL3uc, 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). 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'-GAGCGCGCAGAATACGACTCACTATAGGGCAACCATG-3') and LGT10L (5'-AGCAAGTTCAGCCTGGTTAAG-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
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 $f_i$, 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, was then computed using relative entropy:

$$TSPS_i = \sum_j f_i \log f_i / (q_i)$$

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 of 0 would be reported for TFs expressed uniformly across all tissues, while a maximal TSPS of 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 across 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 S3).

### 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 across 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 (Table S4). The threshold chosen for classifying TFs as tissue “specifiers” (TSPS $\geq 1$) was based on the observed bimodal distribution of expression across 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 S3).

### Data and Analysis Results

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

### 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 (NEXT) 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


