doi:10.1093/nar/gks584

Global identification of transcriptional regulators of pluripotency and differentiation in embryonic stem cells

Kyoung-Jae Won[1,2], Zheng Xu[3], Xian Zhang[2], John W. Whitaker[2], Robert Shoemaker[2], Bing Ren[4], Yang Xu[3] and Wei Wang[2,*]

[1]Department of Genetics, The Institute for Diabetes, Obesity and Metabolism, Perelman School of Medicine at the University of Pennsylvania, 3400 Civic Center Blvd., Philadelphia, PA 19104, [2]Department of Chemistry & Biochemistry, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0359, [3]Division of Biological Sciences, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093 and [4]Ludwig Institute for Cancer Research and Department of Cellular and Molecular Medicine, UCSD School of Medicine, 9500 Gilman Drive, La Jolla, CA 92093, USA

Received June 28, 2011; Revised May 18, 2012; Accepted May 22, 2012

ABSTRACT

Human embryonic stem cells (hESCs) hold great promise for regenerative medicine because they can undergo unlimited self-renewal and retain the capability to differentiate into all cell types in the body. Although numerous genes/proteins such as Oct4 and Gata6 have been identified to play critical regulatory roles in self-renewal and differentiation of hESC, the majority of the regulators in these cellular processes and more importantly how these regulators co-operate with each other and/or with epigenetic modifications are still largely unknown. We propose here a systematic approach to integrate genomic and epigenomic data for identification of direct regulatory interactions. This approach allows reconstruction of cell-type-specific transcription networks in embryonic stem cells (ESCs) and fibroblasts at an unprecedented scale. Many links in the reconstructed networks coincide with known regulatory interactions or literature evidence. Systems-level analyses of these networks not only uncover novel regulators for pluripotency and differentiation, but also reveal extensive interplays between transcription factor binding and epigenetic modifications. Especially, we observed poised enhancers characterized by both active (H3K4me1) and repressive (H3K27me3) histone marks that contain enriched Oct4- and Suz12-binding sites. The success of such a systems biology approach is further supported by experimental validation of the predicted interactions.

INTRODUCTION

The capability of human embryonic stem cells (hESCs) to undergo unlimited self-renewal and retain the pluripotency to differentiate into all cell lineages in the body has raised great hope for developing cell replacement therapy. Although a handful of regulators have been identified to regulate self-renewal and differentiation of hESC, the underlying mechanisms have not been fully understood and additional regulators are still to be uncovered. High-throughput screenings of genes important for maintaining pluripotency in hESC reveal hundreds of potential regulators (1). Distinguishing direct from indirect regulations in such screenings as well as illustrating the mechanisms of these regulators are the immediate challenges.

In this study, we focus on identification of transcription factors (TFs) that may play crucial roles in regulating self-renewal and differentiation of ESCs. The functions of a TF are largely conveyed by its target genes. Despite the availability of complete genome sequences for many organisms, genome-wide identification of TF direct targets and assembling these regulatory interactions into a functional network in mammals remain a challenge (2). chromatin immunoprecipitation (ChIP)-based technologies have been exploited to determine binding sites of numerous TFs in higher organisms (3,4). However, this approach is hindered by the limited

*To whom correspondence should be addressed. Tel: +1 858 822 4240; Fax: +1 858 822 4236; Email: wei-wang@ucsd.edu

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

availability of suitable antibodies and cell types for analysis. Additionally, not all TF-binding sites correspond to functional consequences. In parallel, many computational methods have also been developed to infer transcription networks from gene expression, TF binding data or integration of various types of data (2). When applied to mammalian genomes, these methods often suffer from inability to distinguish direct from indirect target genes or to identify the context-dependent activities of the transcription network.

Recent studies reveal that functional elements such as promoters and enhancers are associated with characteristic chromatin signatures (5–7). Genome-wide maps of chromatin modification states have led to identification of such elements in the human genome (6,8). More recently, DNA methylomes have been mapped at base resolution in multiple cell types (9). These epigenomic data are context dependent and reflect the functional state of the cell. Integrating epigenomic and genomic information to identify transcription factor-binding sites (TFBSs) (10,11) may allow one to determine cell-type-specific transcription networks. In this study, we demonstrated the success of this approach in reconstructing cell-type-specific transcription networks in pluripotent cells (human and mouse ESC, hESC and mESC) and lineage committed (human fetal lung fibroblast, hFLF) cell type. Genome-wide identification of TFBSs allowed reconstruction of these networks (each consisting of >11,000 genes and >500,000 interactions) at an unprecedented scale.

We conducted systems-level analyses of these networks that revealed regulators of pluripotency or differentiation, and illustrated how they might cooperate with the master regulators (Oct4, Nanog and Sox2) in ESCs. Gene expression changes of these predicted regulators upon knockdown of Oct4 in hESC confirmed their functional roles. Furthermore, these networks also facilitated investigation of the interplay between TF binding and epigenetic modifications. Especially, we observed poised enhancers marked by both active (H3K4me1) and repressive (H3K27me3) histone marks that contain enriched Oct4- and Suz12-binding sites in hESC.

MATERIALS AND METHODS

Dataset

Epigenomic data were available in the H1 human embryonic stem cells (hESC), human fetal lung fibroblast IMR90 cells (hFLF) (12,13) and V6.5 mouse embryonic stem cells (mESC) (14,15). In human (H1 and IMR90), we used 11 histone modification marks and DNA methylation (mCG, mCHG and mCHH, where H = A, C or T) (13) data generated by the San Diego Epigenome Center. The 11 histone marks include H2BK5ac, H3K4me1/2/3, H3K9ac, H3K9me3, H3K18ac, H3K27ac, H3K27me3, H3K36me3 and H4K5ac. Gene expressions in H1 and IMR90 were measured by RNA-seq (13). In mESC, we used 8 histone marks (pan-H3, H3K4me1–3, H3K9me3, H4K20me3, H3K27me3 and H3K36me3) but no DNA methylation data were available (14,15). Gene expressions in mESCs were taken from microarray data in (14) and (16).

Assessment of the reconstructed networks using literature

We developed an automatic literature mining tool called STAR miner to evaluate the entire network (see Supplementary Data). STAR miner considers not only the co-occurrence of a TF and its target in the same paper, but also the explicit causal regulation between the entities. For example, an edge from Oct4 to Nanog is considered to be supported by literature only when statements like 'Oct4 activates Nanog' or 'Nanog is regulated by Oct4' are found, suggesting more reliable retrieval of regulatory interactions than those methods that consider only co-occurrence of two entities (17).

Oct4 knockdown

The short hairpin RNA (shRNA)-knockdown vector was made based on the episomal vector pCAGIPuro (18). Briefly, the enhanced green fluorescent protein (EGFP) was cloned into pCAGIPuro, the expression of which is driven by the chick β-actin (CAG) enhancer/promoter that is highly active in hES cells (19). A H1 promoter followed by multiple cloning sites was then introduced. The shRNAs were inserted into the vector using the BglII and SalI sites. The target sequence for knockdown of human Oct4 (NM_002701) was GGATGTGGTCCGA GTGTGG. We also constructed a universal negative control shRNA vector using the sequence, ACTACCGT TGTTATAGGTG, from pSilencer^{TM} 4.1-CMV Expression Vectors (Ambion).

The hES cell line HUES9 was cultured under feeder-free condition using hESC-qualified Matrix (BD Biosciences) and mTeSR^{TM}1 defined medium (Stem Cell Technologies). Cells were passaged using TrypLE (Invitrogen). Negative control shRNA or the Oct4 shRNA constructs were transfected into HUES9 cells using electroporation. Puromycin (0.3 µg/ml) was added to the media the next day to selectively enrich the shRNA-expressing cells. The hES cell line H1 was cultured and transfected the same way as the HUES9 line, except that a rho-associated kinase inhibitor Y27632 (Stemgent) was added to the media after passaging or transfection for 24 hours.

Whole cell lysates were subject to electrophoresis on sodium dodecyl sulfate polyacrylamide gel electrophoresis and transferred to Immun-blot PVDF membrane (Bio-Rad). The antibodies used were Oct4 (Santa Cruz Biotechnology), Nanog (Abcam), Sox2 (Millipore) and α-Tubulin (Sigma). The blots were developed with the SuperSignal system (Thermo Scientific).

Total RNA was isolated using the RNeasy kit (Qiagen). cDNA was synthesized using SuperScript reverse transcriptase III (Invitrogen). qPCR assay was performed using an Applied Biosystems Prism7000 Sequence Detection System. The levels of Oct4 and Nanog were measured using Taqman gene expression assays and Ubiquitin C (UBC) was used as the endogenous control. The levels of other genes were assayed using SYBR Green PCR Master mix (Roche), and Glyceraldehyde 3-phosphate dehydrogenase was used as the endogenous control.

For the Oct4 intracellular staining, cells were fixed in 4% paraformaldehyde, permeabilized with 1% Triton X-100 and blocked in PBS-5% fetal calf serum. Cells were then incubated with 1:100 dilution of an anti-Oct4 antibody (Abcam), Phycoerythrin followed by staining with a Cy3-conjugated donkey-anti-rabbit IgG antibody (Jackson ImmunoResearch). For staining of surface antigens, cells were incubated with PE-conjugated antibodies against stage-specific embryonic antigens (SSEA)-3, Tra-1-60 and Tra-1-81 (Becton Dickison) for 20 minutes in the dark. All samples were analysed on a LSRII flow cytometer (Becton Dickison). Total RNA from control shRNA- and knockdown shRNA-transfected cells were sent to Seqwright for microarray experiment. The Human U133 Plus 2.0 array (Affymetrix) was used.

RESULTS

A systematic approach to reconstructing transcription network

To conduct a comprehensive search for new regulators of self-renewal and differentiation of ESC, we first integrate epigenomic and genomic data to reconstruct transcription networks in a cell-type-specific manner (Figure 1), which consists of the following steps.

Step 1: Genome-wide prediction of promoters and enhancers

Based on the characteristic spatial patterns of chromatin modifications (Supplementary Table S1 and Figure S1) associated with promoters and enhancers, we used Chromia (8) to predict at least 12,000 promoters and 17,000 enhancers in any of the three cell types (Supplementary Figure S2 and Table S2). Presumably, these predicted promoters/enhancers are actively regulated as they are marked by open chromatin.

Step 2: Predicting TFBSs at a genomic scale

Encouraged by the success of Chromia on predicting TFBSs in mESC by incorporating chromatin modification information (11), we searched for TFBSs by scanning 732 known human/mouse motifs in the predicted promoters and enhancers. Additionally, we penalized a motif score if the sequence contained a methylated cytosine(13). In hESC we found 130 enriched (hypergeometric P-value $< 10^{-5}$) motifs in promoters and 161 in enhancers (P-value $< 10^{-5}$) including the master regulators in ESC (Oct4, Sox2 and Nanog).

Step 3: Reconstructing transcription network

Given the genome-wide TFBSs, one needs to assign TFs to their target genes to reconstruct a transcription network. We hypothesized that (i) a gene is regulated by a TF bound to its promoter and (ii) all genes with an actively regulated promoter, as predicted in Step 1, that reside in the same CTCF block of a predicted enhancer can be regulated by TFs bound to the enhancer. Predicted enhancers that did not co-localize with any predicted promoters within a CTCF block were discarded. Such an enhancer-promoter assignment is not perfect but consistent with the recent finding of transcriptional domain in which boundaries of these domains coincide with CTCF-binding sites suggesting the validity of CTCF block (20). On average an enhancer was mapped to 1.7, 1.4 and 1.6 genes in hESC, hFLF and mESC, respectively. False positive enhancer-target assignment can be further reduced when additional data such as HiC (21) or chromatin interaction analysis by paired end tag sequencing (ChIA-PET) (22) become available in these cells. In addition, we also incorporated TF ChIP-seq data (Oct4, Nanog and Sox2) in hESC (12), and 12 TFs in mESC ((Supplementary Table S1) (23)) into the network. To eliminate nonfunctional binding sites in the ChIP-Seq data, we only included the binding peaks supported by the predicted promoters or enhancers (see Supplementary Data).

Transcription networks in hESC, hFLF and mESC

We reconstructed transcription networks, referred as hESnet, hFLFnet and mESnet, respectively, consisting of presumably direct regulatory interactions in hESC (H1), hFLF (IMR90) and mESC (V6.5) (Supplementary Table S3). These are the largest transcription networks reconstructed to date in these cells. For example, the hESC network (hESnet) is composed of 786,250 edges and 13,777 nodes including 174 TFs that are densely interconnected (Figure 2 and Supplementary Table S3). Compared with the yeast network (24), the hESnet has a high clustering coefficient (0.731 compared with 0.189 in yeast). We also reconstructed a conserved ESC network (ESnet) using the intersection between hESnet and mESnet, which represents the evolutionarily conserved regulatory circuitry in the ESC. Note that the corresponding TFBSs are not necessarily aligned in the human and mouse genomes, which avoids errors introduced by the loss/gain/shuffling of the TFBSs during evolution. Interestingly, the ESnet contains more edges resulted from TFs binding to promoters than to enhancers (Supplementary Table S3).

Using the ES-specific gene list obtained from (25), we checked the regulatory relationship among the genes in hESnet (Figure 2 and Supplementary Figure S3). In this sub-network, the key master regulators (Oct4, Sox2 and Nanog) as well as Foxo1, Sox21, Foxm1, HMGA1 and Sox4 are hubs (Supplementary Figure S3). There are notable differences between hESnet and mESnet. For example, Prdm14, a PR domain-containing protein, is highly expressed in hESCs (26) and regulates key pluripotency genes (1), but shows low expression in mES and mouse embryonic germ cells (27). Consistently, Oct4, Sox2 and Nanog regulate Prdm14 in hESnet but not in mESnet. Indeed, Prdm14 was affected by knocking down Oct4 in hESC (Supporting website), but not affected by either Oct4 or Nanog knockdown in the mESC (28), which was confirmed by another independent study that Prdm14 shares many target regions with Nanog and Oct4 but not affected by Oct4 knockdown (29). This example illustrates the advantage of using the context-dependent epigenomic data to identify regulatory interactions specific to species or cell type.

Assessment of the reconstructed networks

Comparison with the networks in public databases

We first compared the predicted regulatory interactions with two networks in the public databases. The first

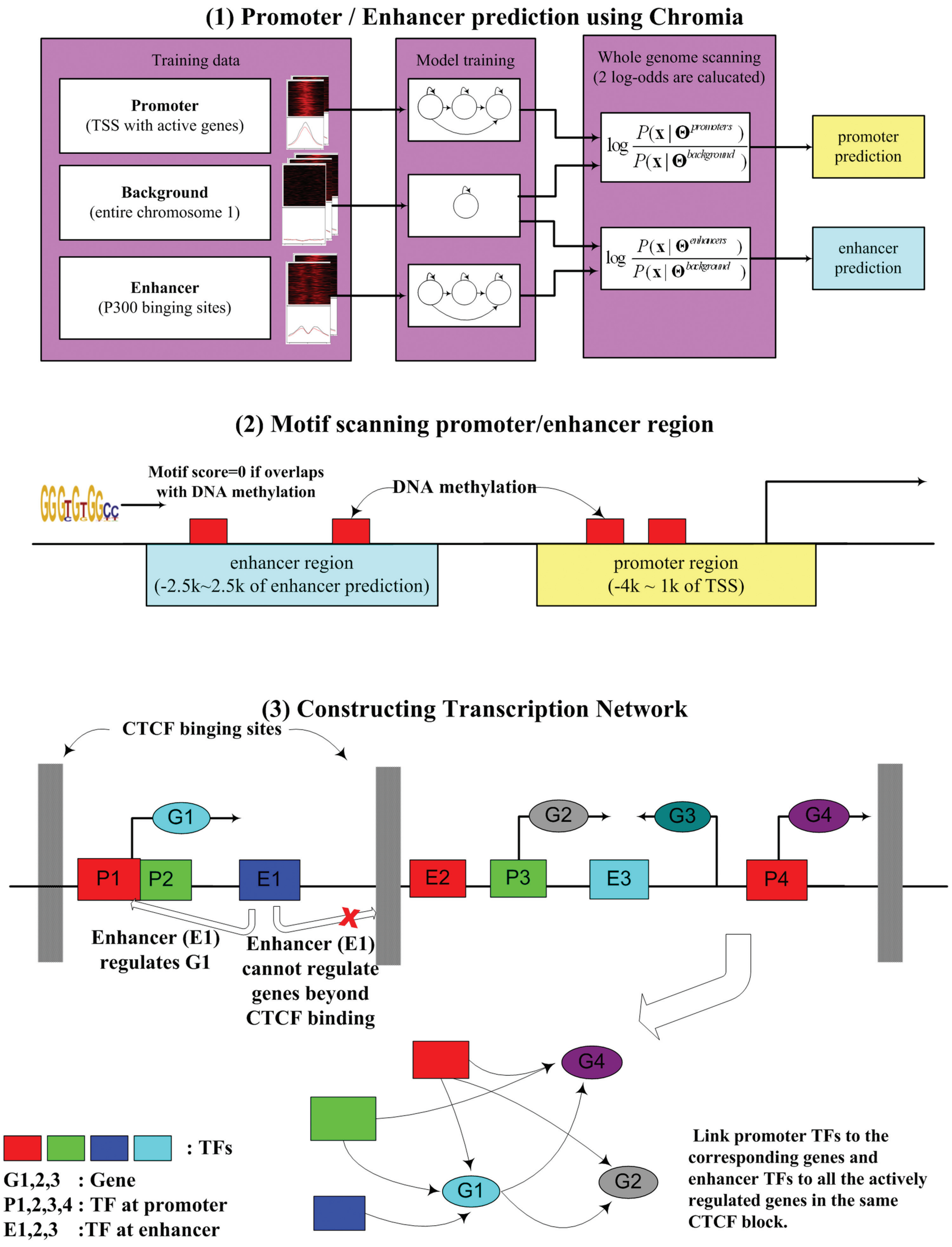
(1) Promoter / Enhancer prediction using Chromia
Training data
Promoter
(TSS with active genes)
Background
(entire chromosome 1)
Enhancer
(P300 binging sites)
Model training
Whole genome scanning
(2 log-odds are calucated)
$\log \frac{P(\mathbf{x}|\Theta^{promoters})}{P(\mathbf{x}|\Theta^{background})}$
$\log \frac{P(\mathbf{x}|\Theta^{enhancers})}{P(\mathbf{x}|\Theta^{background})}$
promoter prediction
enhancer prediction
(2) Motif scanning promoter/enhancer region
Motif score=0 if overlaps with DNA methylation
DNA methylation
enhancer region
(-2.5k~2.5k of enhancer prediction)
promoter region
(-4k ~ 1k of TSS)
(3) Constructing Transcription Network
CTCF binging sites
G1
G2
G3
G4
P1
P2
E1
E2
P3
E3
P4
Enhancer (E1) regulates G1
Enhancer (E1) cannot regulate genes beyond CTCF binding
: TFs
G1,2,3 : Gene
P1,2,3,4 : TF at promoter
E1,2,3 :TF at enhancer
Link promoter TFs to the corresponding genes and enhancer TFs to all the actively regulated genes in the same CTCF block.

Figure 1. Reconstruction of transcription network using genomic and epigenomic information. (1) Genome-wide prediction of promoters and enhancers based on their chromatin signatures using Chromia (8). (2) Prediction of TF-binding sites (TFBSs). Motif scores for each TF were calculated using a sliding window in the predicted promoters and enhancers. The motif score of a window containing a methylated cytosine was set to 0. We only considered the regions of −4∼+1 kbps around the predicted transcription start sites (TSSs) and −2.5∼+2.5 kbps around the predicted enhancers. (3) Direct target genes of each TF were assigned based on their proximity to promoters or enhancers located within the same CTCF block. A TF binding in the promoter of a gene was considered to regulate the gene. A TF bound in enhancers was linked to all the actively regulated genes (with a predicted promoter) within the same CTCF block. In this example, E1 regulates G1, but not G2 or G4 because E1 is insulated from G2 and G4 by CTCF. Gene G1 encodes a TF that binds to E3 and it thus regulates G2 and G4. G3 is not regulated by any TF because it is not actively regulated (no predicted promoter). Finally, all regulatory interactions were assembled into a transcription network.

network is the integrated stem cell molecular interaction database (iScMiD) (30) literature-based network that is reconstructed based on 271 publications and contains both cell-signalling and TF-gene regulatory links (Supplementary Table S4). A caveat of using the iScMid database to evaluate the predicted network is that the iScMid network includes protein–protein interactions (e.g. Nanog-MTA1) and indirect transcriptional regulation (e.g. Sox15-CTGF), which are not expected to be detected by our method. Especially, many Oct4 bindings in iScMid are indeed protein–protein interactions. The overlap between the iScMid literature network and the ESnet is very significant (P-value $= 2.0 \times 10^{-5}$). The second network is embryonic stem cells atlas of pluripotency evidence (ESCAPE) ChIP network (http://www.maayanlab.net/ESCAPE) that is assembled from ChIP-seq experiments of 12 ES-related TFs in mESC (Supplementary Table S5). For the TFs whose targets

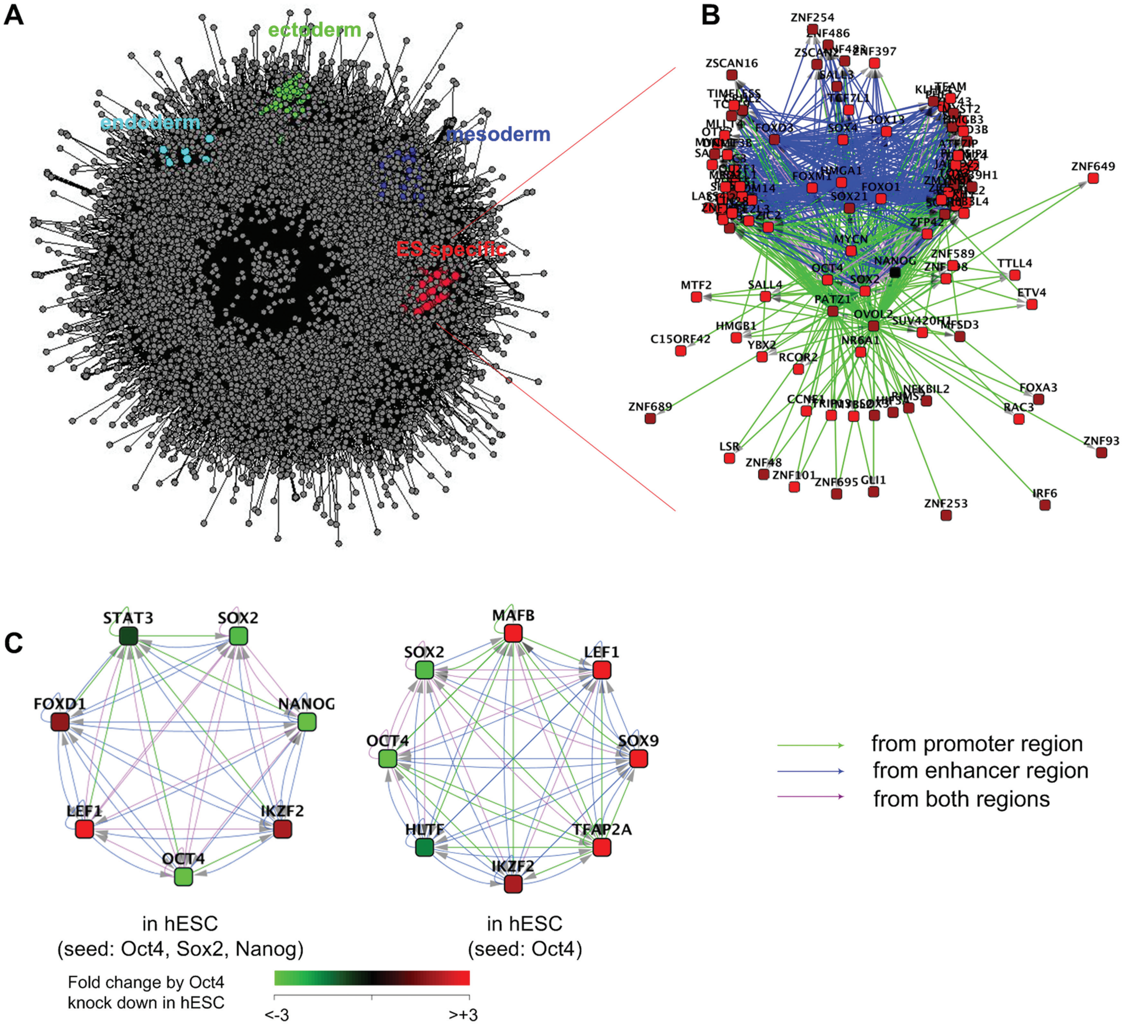

Figure 2. (**A**) The hESnet (ESC and lineage specific genes are highlighted). (**B**) The sub-network composed of ESC-specific TFs (25). (**C**) The largest clique found in hESnet with different seeds.

were only predicted from computation in mESnet, the/' overlap between the predicted and the ChIP-seq the ChIP-seq targets was extremely significant (the least significant P-value was still 4.1×10^{-17}). Since the ESCAPE ChIP network consists of only small fraction of binding, such a significant overlap with the predicted network provides a solid validation of our network.

Literature evidence found by a literature mining tool

To have a large-scale assessment of the predicted networks using literature, we developed STAR miner that automatically retrieves transcriptional regulatory relationships from the published papers (see Supplementary Data, Figure S4). Note that this method does not only consider co-occurrence of a TF and its target in the same paper, but also explicitly considers the causal regulation between the entities. Because TFs are often better studied and have more literature evidence than the other genes, we specifically examined the TF network, which consists of only TF nodes. We observed impressive positive prediction values (PPVs) of 35%, 29% and 41% for hESnet, mESnet and ESnet, respectively. If TFs with low gene expression levels in ESCs were removed, the PPVs in these networks became even higher (38%, 34% and 44%) (Supplementary Figure S7). Because many regulatory interactions discovered in this study might be new, it is not surprising that the PPVs for the entire network were not as high as those for the TF network and the predicted regulatory interactions would be useful in guiding further experimental investigations (The literature evidence is available at http://wanglab.ucsd.edu/star/hESnet/index.jsp).

Oct4 knockdown experiment in hESC

To identify functional targets of Oct4, we silenced Oct4 using an episomal vector-based shRNA knockdown vector and measured gene expression changes on day 3, 5 and 7 using microarray (Figure 3). The microarray measurements were validated using QRT-PCR and the correlation was >0.95 for all 3 days (Supplementary Table S8). When a cutoff of fold change compared with control is 4, the predicted targets include 209 out of 1877 RNAi-affected genes (11.1%), which indicate likely direct targets of Oct4 (Figure 3g). This percentage is comparable

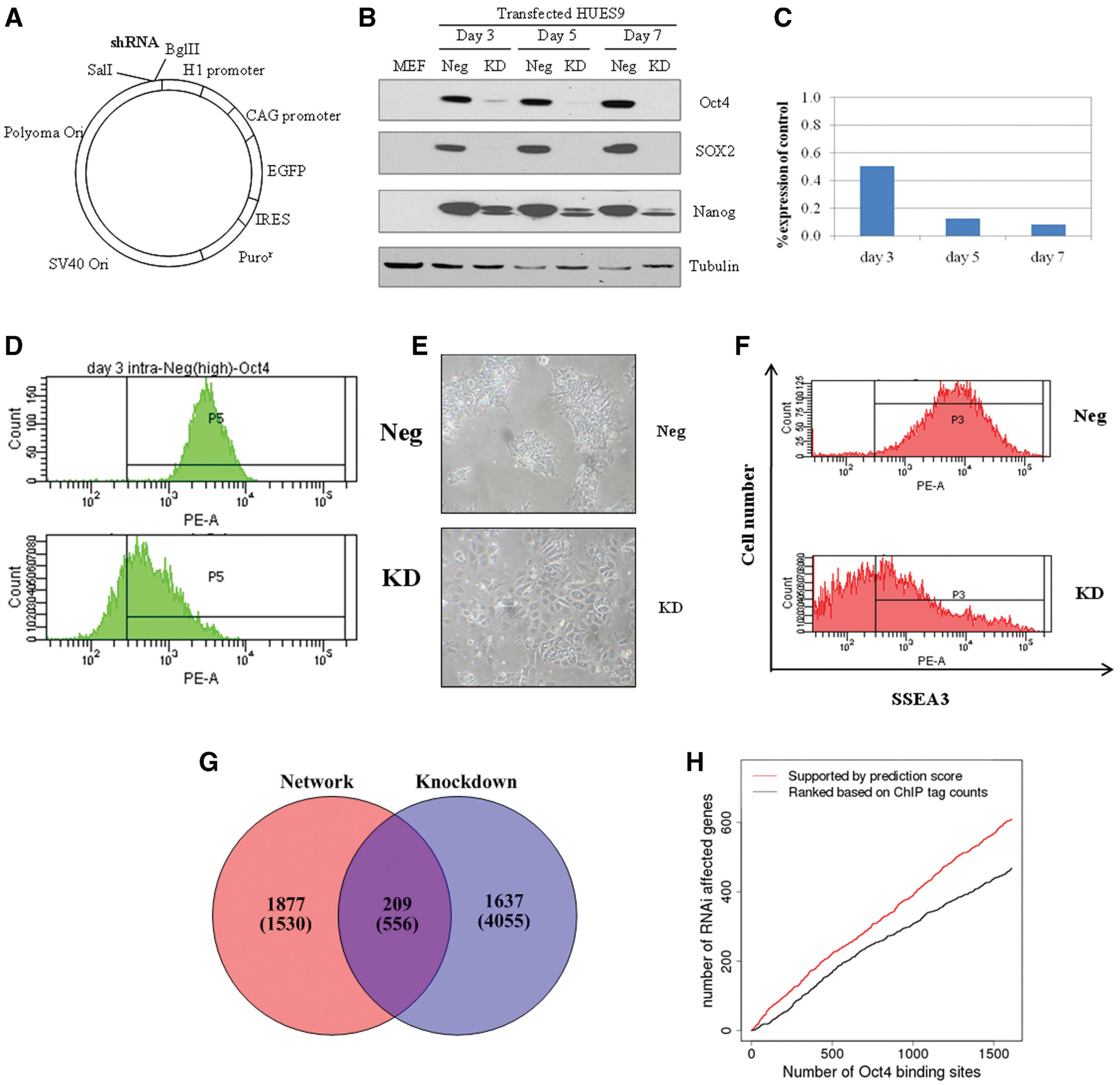

Figure 3. Silencing of Oct4 in HUES9. (**A**) The episomal shRNA vector. The negative control or the Oct4 knockdown shRNA sequence was inserted into the BglII/SalI sites downstream of the H1 promoter. (**B–F**) Silencing of Oct4 in HUES9 cells by the Oct4 shRNA construct. HUES9 cells growing under feeder-free condition were transfected with either the negative control (Neg) or the Oct4 knockdown (KD) construct. Subjected to puromycin selection after 24 hours, cells were harvested at day 3, 5 and 7 post transfection. (B) Western blot analysis using an Oct4-specific antibody. (C) Total RNAs from day 3, 5 and 7 samples were reverse-transcribed and qPCR experiments were performed using a Taqman probe for Oct4. (D) FACS histogram of intracellular Oct4 level in GFP-positive live cells from Neg- or KD-transfected HUES9 at day 3. (E) Phase contrast microscopy of cells transfected with Neg or KD construct at day 5. (F) FACS histogram of SSEA-3 expression on the surface of live cells from Neg- or KD-transfected HUES9 at day 5. (**G**) Overlap between the Oct4 targets in the hESnet and the genes affected by knocking down Oct4 using a fold change cutoff of 4 or 2 (in parentheses). (**H**) Epigenetic information improved identification of functional targets. Among the 6288 Oct4 ChIP binding peaks, 1611 were close to promoter/enhancer predictions. We compared the 1611 RNAi-affected genes with those assigned to the top 1611 ChIP binding peaks (the union of affected genes on day 3, 5 and 7, and the fold change cutoff = 2).

with the overlap between Oct4 ChIP-seq peaks and genes affected by Oct4 knockdown in mESC (11,23,28). Compared with the results of using ChIP-seq peaks (genes within the same CTCF blocks assigned to an Oct4 binding peak), the epigenetic information could consistently improve the prediction accuracy (Figure 3h).

Systems-level analyses of the networks

Identification of new regulators of pluripotency and differentiation

To uncover new regulators of pluripotency and differentiation of ESC, we focused on the TFs that form reciprocal regulations with the known regulators such as Oct4, Sox2 and Nanog. We searched for cliques of TFs in the networks using a heuristic searching algorithm (see Supplementary Data). We defined a clique as a set of nodes that are fully connected by bi-directional edges.

Using Oct4 as a seed, we found 24 unique cliques in hESnet. We ranked TFs based on their occurrences in these cliques. Tfap2a, Lef1, Mafb, Ikzf2, and Stat3 form cliques with Oct4 even more frequently than Sox2 and Nanog (Supplementary Table S9). Consistently, the TFs that more frequently form cliques with Oct4 tend to be more sensitive to Oct4 silencing. Except STAT3, all TFs appearing ≥5 cliques had an absolute fold change >2 in Oct4-knockdown experiments (an example shown in Figure 2c). These evidence suggest functional roles of these TFs in pluripotency maintenance or differentiation.

Intrigued by this analysis, we investigated all TFs that form cliques with different combinations of the three key regulators (Oct4, Sox2 and Nanog) (Supplementary Table S10) and found four unique cliques that all contain Lef1 and Ikzf2. The TFs that form the largest clique in hESC are known to play crucial roles in pluripotency maintenance (STAT3) or differentiation (IKZF2, LEF1, FOXD1 and FOXJ1). The distinct cliques for Oct4, Sox2 and Nanog suggest their own unique roles besides their shared functions. (Supplementary Table S11 and Figure S5). This observation is reminiscent of the overlapping but distinguishable binding sites of these TFs in both human and mouse (23,31) as well as a previous study that showed Nanog and Oct4 function in parallel pathways (32).

Functional cooperation between TFs revealed by common target genes

To investigate the functional cooperation between TFs, we clustered TFs using their normalized common targets (see Supplementary Data) and several major clusters were found (Figure 4 and Supplementary Figure S6). For example, GATA4 and Foxa2 in hESnet, and Runx2 and VDR in hFLFnet share many targets even though they recognize distinct motifs (Figure 4, motifs in Supplementary Table S12). GATA and Fox proteins are part of the evolutionarily conserved network regulating endoderm development (33) and play roles in the hepatic development (34). Their common targets are indeed involved in embryonic morphogenesis (P-value = 8.0×10^{-10}) and pattern specification process (P-value = 1.2×10^{-6}), which confirmed their roles in development. Runx2 and VDR, key TFs for osteoblastogenesis, form a complex to bind to DNA during skeletal development (35), which is consistent with the functions of their common targets: cell death (P-value = 2.9×10^{-10}), blood vessel development (P-value = 3.4×10^{-7}) and skeletal system development (P-value = 2.6×10^{-4}).

Network rewiring during differentiation

Because hESC and hFLF are at the two ends of differentiation, the networks in the two cell types allow us investigate how the transcription network is rewired during differentiation. We first found the gain and loss of edges caused by enhancer activity change in the hESC and hFLF. We only examined genes that are actively regulated (with a predicted promoter in both cell types) and also differentially expressed, i.e. high expression in one cell type and low in another (see Supplementary Data). Among the 241 genes up-regulated in hFLF from hESC, 109 genes show changes of enhancer activity. Among the 301 down-regulated genes in hFLF from hESC, 98 genes show changes in enhancer activity. For example, *Adamts19* has a low gene expression in hFLF but high in hESC (Supplementary Figure S8). The chromatin patterns at its promoter are quite similar between hESC and hFLF. However, an enhancer was found upstream of the gene only in hESC but not in hFLF (Supplementary Figure S8).

DNA methylation often prevents TF binding to DNA (36,37). To investigate its effect on transcriptional regulation, we checked the 19 TFs that are highly expressed in both hESC and hFLF, and artificially applied the DNA methylation profile of the hFLF to hESC in the promoters of their target genes (Supplementary Figure S9): among 241 genes expressed highly in hFLF and lowly in hESC, 159 edges would be added between the 19 TFs and 67 genes; among 301 genes expressed highly in hESC and lowly in hFLF, 172 edges in hESnet formed between 19 TFs and 71 genes would be removed by the hFLF DNA methylation. This observation confirms the repressive role

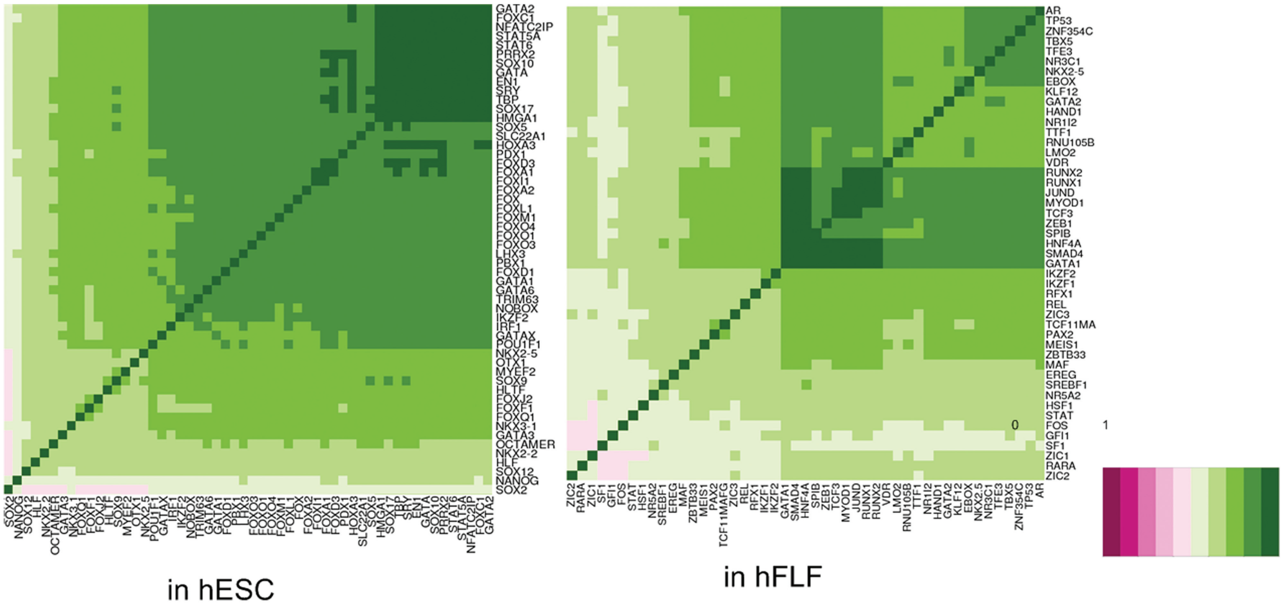

Figure 4. Clustering TFs based on their common targets in hESnet and hFLFnet (complete figures in Supplementary Figure S3).

of DNA methylation in promoter regions. Gene ontology analysis shows that large portion of these genes is related to membrane (Supplementary Figure S10), which indicates a broad role of DNA methylation in regulating membrane proteins during differentiation (38). Furthermore, we found a majority of the genes up- or down-regulated in hFLF contain promoters that lack a TATA box, regardless of the CpG content (Supplementary Figure S11). In summary, either the change of enhancer activity or DNA methylation in promoters can explain 249 out of 542 (46%) of differentially expressed genes in hESC and hFLF (Supplementary Table S13).

The cell-type-specific networks also allow identification of putative lineage-specific TFs. We collected 80 mesoderm-, 257 ectoderm- and 19 endoderm-specific genes (see Supplementary Data) and examined the change of edges between these genes and their regulators in hESnet and hFLFnet. As the hFLF IMR90 cell is endodermal, we found increase of regulatory interactions from hESC to hFLF between six TFs and the endoderm-specific genes, suggesting possible roles of these TFs in endoderm differentiation (Figure 5). Indeed, Smad4 has a known function in endodermal cell migration observed in the mouse development (Supplementary Table S14). Interestingly, we observed a set of TFs regulating less mesoderm-specific genes in hFLF than in hESC, including Runx1, Nkx2-5, Tcf3, Gata1, Ikzf2 (Figure 5), all of which are supported by literature (Supplementary Table S14). Similarly, SETD2 regulates less ectoderm-specific genes in hFLF than in hESC.

Interplay between TF binding and epigenetic modifications in hESC

Determination of TF-binding sites provides an opportunity to investigate how their bindings interact with epigenetic modifications. We focused on enhancers in hESC and categorized them based on their distinct epigenetic signals (Supplementary Figure S12 and Figure 6). Enrichment of Oct4, Sox2 and Nanog binding in the hESnet was examined in each cluster. All three key regulators bind preferentially to cluster 1 (Table 1, Supplementary Tables S15, S16 and S10). In contrast, only Sox2 but neither Oct4 nor Nanog prefers cluster 4, which lacks the dip of mCG at the enhancer centers as shown in cluster 1. It is noteworthy that enhancer predictions were made only using histone modifications without considering DNA methylation. This observation may suggest that CG and non-CG methylation may play distinct roles in interplaying with histone modifications or protein binding.

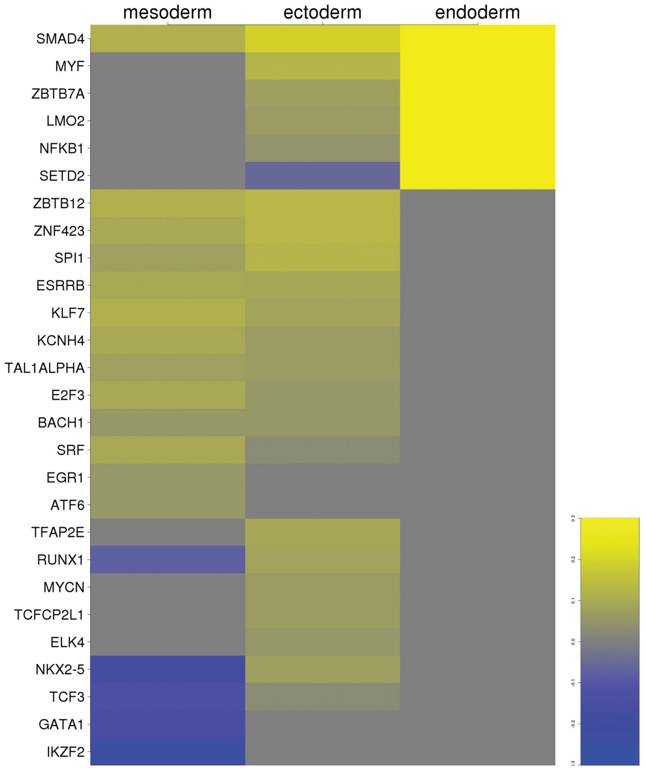

Figure 5. Lineage specificity of TFs. The normalized TF target differences (hFLFnet – hESnet) were calculated as the metric to cluster TFs using a hierarchical clustering algorithm.

Cluster 2, 6 and 7 also present biased binding of these TFs (Table 1). Especially, only Oct4 binding is enriched in cluster 7 and 42% of the Oct4 targets in this cluster are confirmed by the knockdown experiments (Table 1, Supplementary Tables S15 and S16). Enriched functions of these Oct4 targets are related to development (Supplementary Data). Cluster 7 shows strong active H3K4me1/2 and repressive H3K27me3 marks, which is reminiscent of the bivalent H3K4me3/H3k27me3 pattern found in promoters (39) and consistent with the poised enhancers (40,41). All the genes within the same CTCF blocks of these enhancers (particularly those actively regulated ones) are expressed significantly lower in hESC (t-test P-value $< 2.2 \times 10^{-16}$) than other clusters (Supplementary Table S15) but much higher than other clusters upon Oct4 knockdown (Figure 6b). Notably, the poised enhancers contain enriched binding sites of Suz12 (42), a subunit of the Polycomb Repressive Complex 2 (PRC2) (Table 1 and Supplementary Table S16), which is consistent with the role of Oct4–Polycomb interaction for lineage specific repression (42,43). Cluster 7 also lacks the characteristic non-CG methylation in hESC. To search for additional TFs that may bind to the bivalent enhancers, we conducted motif enrichment analyses and found additional TFs such as Myc and Znf219 that may interplay with the bivalent enhancers (see Supplementary Data).

DISCUSSION

We demonstrate here the first attempt to identify new regulators of pluripotency and differentiation in ESCs using a systems biology approach. We reconstructed transcription networks using epigenomic data at a genome-wide scale, which revealed many new TFs that likely cooperate with the known master regulators to regulate self-renewal and differentiation. The Oct4-knockdown experiments further supported the possible functional roles of these TFs in ES cells. Regulatory interactions

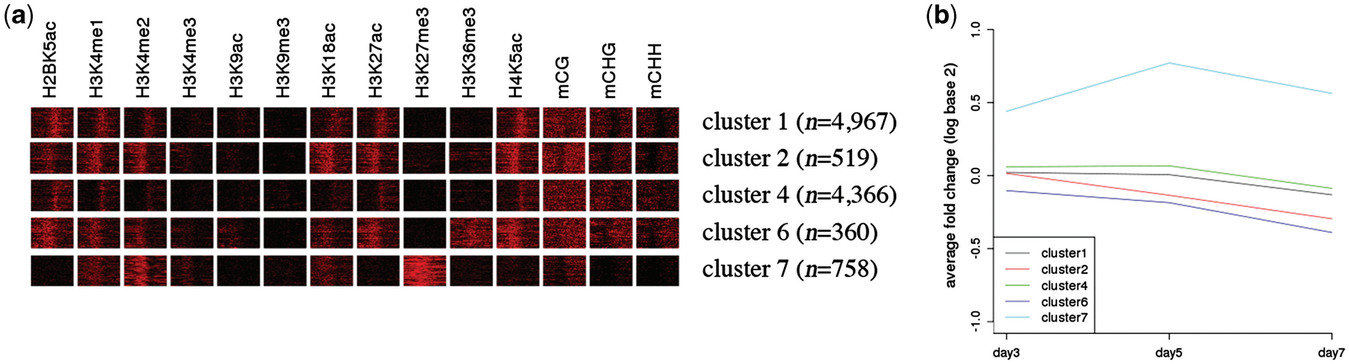

Figure 6. Representative enhancer clusters. (**a**) Epigenetic profiles at hESC enhancers (complete figure in Supplementary Figure S5). (**b**) Average expression change of the genes regulated by the enhancers in each cluster upon Oct4 knockdown.

Table 1. Occurrence of the TF targets in each cluster. Hypergeometric *P*-value was calculated to evaluate the significance of the enrichment

Cluster	Number of enhancers	Number of genes mapped to enhancers	Average RPKMs of genes mapped to enhancers	Oct4 target genes			Sox2 target genes		Nanog target genes		Suz12 binding peaks (%)
				Number of targets	*P*-value	RNAi-affected genes	Number of targets	*P*-value	Number of targets	*P*-value	
1	4967	2717	24.4	142	$\mathbf{2.8 \times 10^{-11}}$	52	1813	$\mathbf{6.3 \times 10^{-320}}$	1681	$\mathbf{2.8 \times 10^{-295}}$	0
2	519	494	58.3	39	$\mathbf{5.15 \times 10^{-6}}$	17	248	0.02	239	$\mathbf{2.1 \times 10^{-4}}$	0
4	4366	2647	20.6	21	1.0	11	1116	$\mathbf{8.9 \times 10^{-3}}$	914	0.97	0
6	360	422	50.4	25	0.01	8	294	$\mathbf{1.3 \times 10^{-29}}$	264	$\mathbf{7.3 \times 10^{-24}}$	0
7	758	667	7.7	33	$\mathbf{7.8 \times 10^{-4}}$	14	119	1	120	1	**27**

RNAi-affected genes are genes whose expression values changed more than 2 fold after Oct4 knockdown on day 3, 5 or 7. For Suz12 binding, the percentages of the enhancers containing the Suz12-binding sites were calculated. The average reads Per kilobase per million mapped read (RPKM) of all genes is 26.2 in hESC and 22.6 in hFLF (complete table in Supplementary Table S15).

suggested by our analyses may shed light on the mechanisms of pluripotency maintenance and lineage-specific differentiation.

The integrative analysis of genomic and epigenomic data revealed interplay between TF binding and epigenetic modifications. Particularly interesting, a set of bivalent enhancers regulating genes poised in stem cells coincide with preferred Oct4 and Suz12 binding. Target genes of these enhancers are up-regulated upon Oct4 knockdown, which further illustrates the cooperation of Oct4 and Suz12 in repressing cell differentiation and maintaining self-renewal.

Inferring transcription networks based on the cell-state-specific epigenetic modifications also provides a unique opportunity to study the dynamic rewiring of transcription networks. Comparison of gene expression and regulatory interactions in different cell types suggests possible mechanisms such as DNA methylation and enhancer activity change, through which the cell-type-specific gene expression is achieved.

There is still a great room for improvement of the method to reconstruct transcription networks. First, the assignment of enhancers to genes is obviously not optimal. The fast accumulation of genomic and epigenetic data in large number of cells provides an unprecedented opportunity to improve the inference of enhancer-promoter regulation, and these additional data can be incorporated into network reconstruction. Second, the predicted enhancers included both active and poised ones. A further distinction between the two groups of enhancers will no doubt further improve the quality of the reconstructed transcription network. Third, we predicted the binding of expressed TFs based on the presence of their motif in the promoters or enhancers, which is an assumption that is not always true because the TF binding may depend on the binding of its co-factors. To completely resolve these issues, both new experimental techniques and computational methods are needed and are indeed being actively pursued.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–17 and Supplementary Figures 1–12.

ACKNOWLEDGEMENTS

The authors thank Ron Stewart and James Thomson for providing the ES cell-specific gene list. K.J.W. and W.W. conceived and designed the experiments. K.J.W. analyzed the data. X.Z. developed the literature mining tool, designed the companion webpage and wrote the corresponding part of the manuscript. Z.X. and Y.X. performed the Oct4 knockdown experiments and contributed the description of the experimental procedure. J.W.W. and

R.S. contributed to data analysis. B.R. provided chromatin modification data and contributed to analysis of the data. W.W. analyzed the data and supervised the whole study. K.J.W., Z.X., X.Z. and W.W. wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

National Institutes of Health (NIH) [R01GM072856 to W.W., U01ES017166 to B.R. and W.W.]; CIRM [RC100148 to Y.X.]. Funding for open access charge: NIH.

Conflict of interest statement. None declared.

REFERENCES

1. Chia,N.Y., Chan,Y.S., Feng,B., Lu,X., Orlov,Y.L., Moreau,D., Kumar,P., Yang,L., Jiang,J., Lau,M.S. *et al.* (2010) A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. *Nature*, **468**, 316–320.
2. Kim,H.D., Shay,T., O'Shea,E.K. and Regev,A. (2009) Transcriptional regulatory circuits: predicting numbers from alphabets. *Science*, **325**, 429–432.
3. Park,P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–680.
4. Farnham,P.J. (2009) Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.*, **10**, 605–616.
5. Hon,G., Ren,B. and Wang,W. (2008) ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. *PLoS Comput. Biol.*, **4**, e1000201.
6. Heintzman,N.D., Stuart,R.K., Hon,G., Fu,Y., Ching,C.W., Hawkins,R.D., Barrera,L.O., Van Calcar,S., Qu,C., Ching,K.A. *et al.* (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, **39**, 311–318.
7. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
8. Won,K.J., Chepelev,I., Ren,B. and Wang,W. (2008) Prediction of regulatory elements in mammalian genomes using chromatin signatures. *BMC Bioinformatics*, **9**, 547.
9. Laird,P.W. (2010) Principles and challenges of genome-wide DNA methylation analysis. *Nat. Rev. Genet.*, **11**, 191–203.
10. Whitington,T., Perkins,A.C. and Bailey,T.L. (2009) High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites. *Nucleic Acids Res.*, **37**, 14–25.
11. Won,K.J., Ren,B. and Wang,W. (2010) Genome-wide prediction of transcription factor binding sites using an integrated model. *Genome Biol.*, **11**, R7.
12. Hawkins,R.D., Hon,G.C., Lee,L.K., Ngo,Q., Lister,R., Pelizzola,M., Edsall,L.E., Kuan,S., Luu,Y., Klugman,S. *et al.* (2010) Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell*, **6**, 479–491.
13. Lister,R., Pelizzola,M., Dowen,R.H., Hawkins,R.D., Hon,G., Tonti-Filippini,J., Nery,J.R., Lee,L., Ye,Z., Ngo,Q.M. *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.
14. Mikkelsen,T.S., Ku,M., Jaffe,D.B., Issac,B., Lieberman,E., Giannoukos,G., Alvarez,P., Brockman,W., Kim,T.K., Koche,R.P. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.
15. Meissner,A., Mikkelsen,T.S., Gu,H., Wernig,M., Hanna,J., Sivachenko,A., Zhang,X., Bernstein,B.E., Nusbaum,C., Jaffe,D.B. *et al.* (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, **454**, 766–770.
16. Ivanova,N., Dobrin,R., Lu,R., Kotenko,I., Levorse,J., DeCoste,C., Schafer,X., Lun,Y. and Lemischka,I.R. (2006) Dissecting self-renewal in stem cells with RNA interference. *Nature*, **442**, 533–538.
17. Jensen,L.J., Kuhn,M., Stark,M., Chaffron,S., Creevey,C., Muller,J., Doerks,T., Julien,P., Roth,A., Simonovic,M. *et al.* (2009) STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, D412–D416.
18. Lin,T., Chao,C., Saito,S., Mazur,S.J., Murphy,M.E., Appella,E. and Xu,Y. (2005) p53 induces differentiation of mouse embryonic stem cells by suppressing Nanog expression. *Nat. Cell Biol.*, **7**, 165–171.
19. Song,H., Chung,S.K. and Xu,Y. (2010) Modeling disease in human ESCs using an efficient BAC-based homologous recombination system. *Cell Stem Cell*, **6**, 80–89.
20. Dixon,J.R., Selvaraj,S., Yue,F., Kim,A., Li,Y., Shen,Y., Hu,M., Liu,J.S. and Ren,B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.
21. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
22. Fullwood,M.J., Liu,M.H., Pan,Y.F., Liu,J., Xu,H., Mohamed,Y.B., Orlov,Y.L., Velkov,S., Ho,A., Mei,P.H. *et al.* (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, **462**, 58–64.
23. Chen,X., Xu,H., Yuan,P., Fang,F., Huss,M., Vega,V.B., Wong,E., Orlov,Y.L., Zhang,W., Jiang,J. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.
24. Yu,H., Zhu,X., Greenbaum,D., Karro,J. and Gerstein,M. (2004) TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. *Nucleic Acids Res.*, **32**, 328–337.
25. Yu,J., Vodyanik,M.A., Smuga-Otto,K., Antosiewicz-Bourget,J., Frane,J.L., Tian,S., Nie,J., Jonsdottir,G.A., Ruotti,V., Stewart,R. *et al.* (2007) Induced pluripotent stem cell lines derived from human somatic cells. *Science*, **318**, 1917–1920.
26. Assou,S., Le Carrour,T., Tondeur,S., Strom,S., Gabelle,A., Marty,S., Nadal,L., Pantesco,V., Reme,T., Hugnot,J.P. *et al.* (2007) A meta-analysis of human embryonic stem cells transcriptome integrated into a web-based expression atlas. *Stem Cells*, **25**, 961–973.
27. Yamaji,M., Seki,Y., Kurimoto,K., Yabuta,Y., Yuasa,M., Shigeta,M., Yamanaka,K., Ohinata,Y. and Saitou,M. (2008) Critical function of Prdm14 for the establishment of the germ cell lineage in mice. *Nat. Genet.*, **40**, 1016–1022.
28. Loh,Y.H., Wu,Q., Chew,J.L., Vega,V.B., Zhang,W., Chen,X., Bourque,G., George,J., Leong,B., Liu,J. *et al.* (2006) The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.*, **38**, 431–440.
29. Ma,Z., Swigut,T., Valouev,A., Rada-Iglesias,A. and Wysocka,J. (2010) Sequence-specific regulator Prdm14 safeguards mouse ESCs from entering extraembryonic endoderm fates. *Nat. Struct. Mol. Biol.*, **18**, 120–127.
30. Macarthur,B.D., Ma'ayan,A. and Lemischka,I.R. (2009) Systems biology of stem cell fate and cellular reprogramming. *Nat. Rev. Mol. Cell. Biol.*, **10**, 672–681.
31. Boyer,L.A., Lee,T.I., Cole,M.F., Johnstone,S.E., Levine,S.S., Zucker,J.P., Guenther,M.G., Kumar,R.M., Murray,H.L., Jenner,R.G. *et al.* (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, **122**, 947–956.
32. Ralston,A. and Rossant,J. (2005) Genetic regulation of stem cell origins in the mouse embryo. *Clin. Genet.*, **68**, 106–112.
33. Zaret,K.S. (2008) Genetic programming of liver and pancreas progenitors: lessons for stem-cell differentiation. *Nat. Rev. Genet.*, **9**, 329–340.
34. Zaret,K.S. (2002) Regulatory phases of early liver development: paradigms of organogenesis. *Nat. Rev. Genet.*, **3**, 499–512.
35. Marcellini,S., Bruna,C., Henriquez,J.P., Albistur,M., Reyes,A.E., Barriga,E.H., Henriquez,B. and Montecino,M. Evolution of the interaction between Runx2 and VDR, two transcription factors involved in osteoblastogenesis. *BMC Evol. Biol.*, **10**, 78.

36. Prendergast,G.C., Lawe,D. and Ziff,E.B. (1991) Association of Myn, the murine homolog of max, with c-Myc stimulates methylation-sensitive DNA binding and ras cotransformation. *Cell*, **65**, 395–407.
37. Comb,M. and Goodman,H.M. (1990) CpG methylation inhibits proenkephalin gene expression and binding of the transcription factor AP-2. *Nucleic Acids Res.*, **18**, 3975–3982.
38. Remus,R., Kanzaki,A., Yawata,A., Wada,H., Nakanishi,H., Sugihara,T., Zeschnigk,M., Zuther,I., Schmitz,B., Naumann,F. *et al.* (2005) Relationships between DNA methylation and expression in erythrocyte membrane protein (band 3, protein 4.2, and beta-spectrin) genes during human erythroid development and differentiation. *Int. J. Hematol.*, **82**, 422–429.
39. Bernstein,B.E., Mikkelsen,T.S., Xie,X., Kamal,M., Huebert,D.J., Cuff,J., Fry,B., Meissner,A., Wernig,M., Plath,K. *et al.* (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, **125**, 315–326.
40. Rada-Iglesias,A., Bajpai,R., Swigut,T., Brugmann,S.A., Flynn,R.A. and Wysocka,J. (2011) A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, **470**, 279–283.
41. Creyghton,M.P., Cheng,A.W., Welstead,G.G., Kooistra,T., Carey,B.W., Steine,E.J., Hanna,J., Lodato,M.A., Frampton,G.M., Sharp,P.A. *et al.* (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. USA*, **107**, 21931–21936.
42. 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.
43. Squazzo,S.L., O'Geen,H., Komashko,V.M., Krig,S.R., Jin,V.X., Jang,S.W., Margueron,R., Reinberg,D., Green,R. and Farnham,P.J. (2006) Suz12 binds to silenced regions of the genome in a cell-type-specific manner. *Genome Res.*, **16**, 890–900.