A gene regulatory network in mouse embryonic stem cells

Qing Zhou*, Hiram Chipperfield†, Douglas A. Melton†, and Wing Hung Wong‡§

*Department of Statistics, University of California, Los Angeles, 8125 Math Science Building, Los Angeles, CA 90095; †Department of Molecular and Cellular Biology, Harvard University, 7 Divinity Avenue, Cambridge, MA 02138; and ‡Departments of Statistics, Health Research and Policy, and Biological Sciences, Stanford University, 390 Serra Mall, Stanford, CA 94305

Edited by Philip P. Green, University of Washington School of Medicine, Seattle, WA, and approved August 21, 2007 (received for review February 5, 2007)

We analyze new and existing expression and transcription factor-binding data to characterize gene regulatory relations in mouse ES cells (ESC). In addition to confirming the key roles of Oct4, Sox2, and Nanog, our analysis identifies several genes, such as Esrrb, Stat3, Tcf7, Sall4, and LRH-1, as statistically significant coregulators. The regulatory interactions among 15 core regulators are used to construct a gene regulatory network in ESC. The network encapsulates extensive cross-regulations among the core regulators, highlights how they may control epigenetic processes, and reveals the surprising roles of nuclear receptors. Our analysis also provides information on the regulation of a large number of putative target genes of the network.

cis-regulatory module | transcriptional regulation

Recent research has established the fundamental roles of several transcription factors (TF), namely Oct4, Sox2, and Nanog, in the self-renewal and pluripotency of mouse ES cells (ESC) (reviewed in ref. 1). In addition to these three "master regulators," a large number of additional TF have been implicated to play a role in ESC biology, including Stat3, Esrrb, Tbx3, Foxd3, LRH-1, Klf4, Myc, P53, and Sall4 (2–8). The emerging picture is that these TF regulate each other and interact with epigenetic control factors to form a large gene regulatory network in ESC. However, the regulatory interactions within this network have not been worked out. Here, we analyze new and published gene expression and location (ChIP-chip/ChIP-PET) data sets to reconstruct a part of this network computationally. We identify collaborating TF that may work with the master regulators to activate gene expression in ESC, as well as collaborating factors that may play a repressive role. Our analyses also predict the cis-regulatory sequences (down to the location of binding sites) that mediate the combinatorial control of these TF on their target genes.

Results

Genes Correlated or Affected by Oct4 Expression. Several lines of evidence point to Oct4 as one of the most important regulators in ESC. Null mutants of Oct4 are not viable and fail to form a functional inner cell mass (9). A twofold increase or decrease of Oct4 levels leads to differentiation into primitive endoderm/mesoderm or trophectoderm, respectively (10). ESC show a very dramatic change in gene expression when subject to RNAi knockdown of Oct4, as compared with RNAi knockdowns of other essential regulators (3). Moreover, Oct4 is one of the four regulators that together can reprogram fibroblast cells to pluripotent cells with ESC-like morphology (6). No other ESC regulator shares these properties. Because a transcriptional target is likely to have expression correlated with its regulator, the first step in our analysis was to identify genes whose expression strongly correlates with that of Oct4. To do this, we used an ES line harboring a GFP reporter driven by the Oct4 distal enhancer (11). FACS based on this reporter allowed the purification of subpopulations of cells according to their Oct4 expression level. Specifically, at various times after the ESC began to differentiate, we performed FACS on the cell cultures and dissociated embryoid bodies (EB) to separate differentiated (GFP−) from undifferentiated cells (GFP+). These samples produced 16 expression profiles, including 3 profiles of undifferentiated ESC (which is of course high in Oct4 expression); 5 profiles from 2-, 4-, and 8-day EB with high Oct4 expression; and 8 profiles from 2-, 4-, 8-, and 15-day EB with low Oct4 expression (Fig. 1*A* and *Materials and Methods*). Consistent with the role of Oct4 as a master regulator, dramatic differences in mRNA expression are detected between the eight Oct4-high and the eight Oct4-low samples. ESC markers, such as Oct4/Pou5f1, Sox2, Nanog, Esrrb, Tcl1, Dppa5, and Utf1, show very high fold changes (>9 in a positive direction), whereas differentiated cell markers, such as Gata4, Gata6, Foxa2, and Bmp2, show very high fold changes in a negative direction (>10) (Fig. 1*B*). The distributions of fold changes are plotted in supporting information (SI) Fig. 4, with more than 2,000 genes showing more than a threefold change when Oct4-high and -low samples were compared. We selected probe sets with a fold change (in either direction) > 2 and P value < 0.05 in a two-sample comparison to obtain 2,359 Oct4-sorted+ probe sets and 2,784 Oct4-sorted− probe sets showing positive and negative fold changes, respectively (the false discovery rate was <44%). Mapping these probe sets to Refseq genes, we obtain 1,325 Oct4-sorted+ genes and 1,440 Oct4-sorted− genes, respectively. See *SI Text* for the details of gene expression analysis.

We compared our data to the Oct4-RNAi data and retinoic acid (RA) induction data reported by Lemischka and colleagues (3). Significant overlaps are seen between our positive fold change genes (Oct4-sorted+) and genes suppressed by Oct4 knockdown (Oct4-Ri− genes) or by RA-induced differentiation (RA− genes) ($P < 10^{-50}$, Fig. 1*C*). Likewise, our negative fold change genes (Oct4-sorted−) overlap significantly with those induced by Oct4 knockdown (Oct4-Ri+) or by RA induction (RA+). Although the overlaps are very significant statistically, the majority of the genes in each set are not identified in the other sets. Thus we pooled the genes in the Oct4-sorted+, Oct4-RNAi−, and RA− gene sets to obtain 2,045 genes (called Oct4+ genes hereafter). See SI Data Sets 1–7 for all the expression profiles and gene sets.

Combining Expression Data and TF-Binding Data. We hypothesized that computational inference of a part of the gene regulatory network in ESC may be achieved by using the gene expression data described above and recently published binding data for Oct4 and Nanog (12), and Phc1 (13). Phc1 is a component of polycomb complex PRC1 found on promoter regions of a large number of genes repressed in ESC (13). We did not include in this analysis the binding data of the other three polycomb complex proteins, Rnf2, Eed, and Suz12 (13), because their expression fold changes (≈2.5)

Author contributions: D.A.M. and W.H.W. designed research; Q.Z. and H.C. performed research; Q.Z. and W.H.W. analyzed data; and Q.Z. and W.H.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Abbreviations: TF, transcription factor; ESC, ES cell; EB, embryoid body; RA, retinoic acid.

§To whom correspondence should be addressed. E-mail: whwong@stanford.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0701014104/DC1.

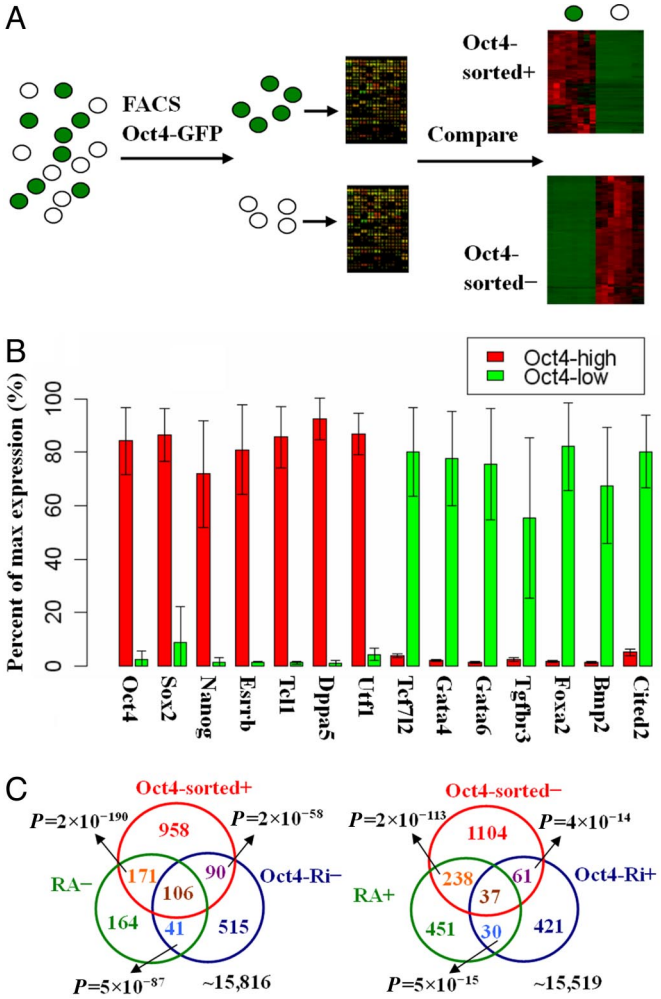

Fig. 1. Overview of the Oct4-sorted series. (*A*) FACS in an ESC to EB differentiation time course followed by expression profiling. The cells were separated by FACS into undifferentiated (GFP+) and differentiated (GFP−) subpopulations, as indicated by green and white solid circles, respectively. Microarray expression profiling was performed on the sorted subpopulations of cells, and we used the resulting data as the basis for identifying Oct4-sorted± genes. (*B*) Expression levels of selected marker genes in the Oct4-high and -low samples as measured by the percentage of maximal expression. The error bars indicate the standard errors of the expression levels. (*C*) Overlaps between the gene sets in the Oct4-sorted, Oct4-RNAi, and RA-induction series. The numbers in the Venn diagrams represent the counts of the contiguous regions (gene sets) where they appear. The *P* values of the overlaps are calculated by the hypergeometric distribution (*SI Text*).

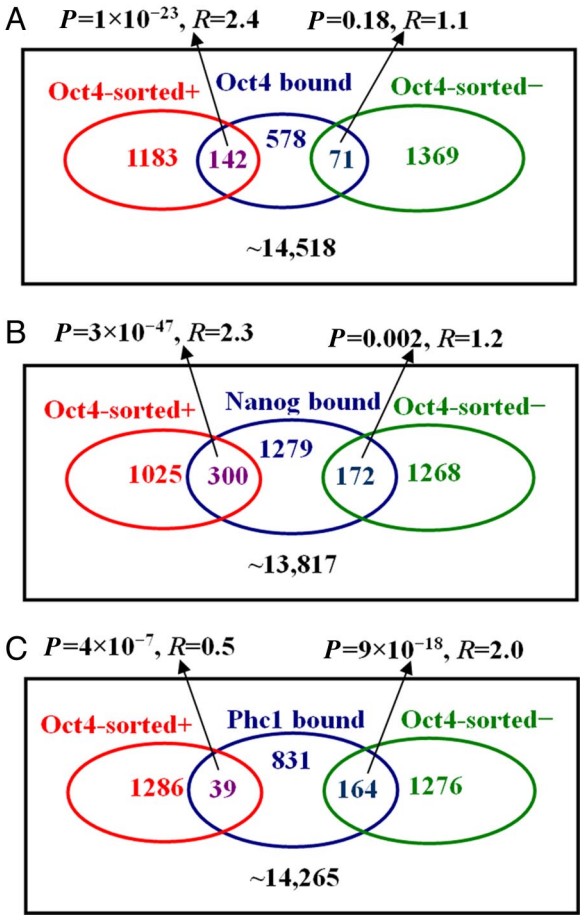

Fig. 2. The overlaps between Oct4-sorted± genes and Oct4-bound (*A*), Nanog-bound (*B*), and Phc1-bound (*C*) genes. The statistical significance (*P* value) is calculated on the basis of the hypergeometric distribution, and the enrichment level *R* is defined as the ratio of the number of observed overlaps over that of expected overlaps. The computations of the *P* value and enrichment level are explained in *SI Text*.

in our Oct4-sorted series were much lower than that of Phc1 (≈5.8). We associated 1,083 Oct4-binding regions, 3,006 Nanog-binding regions, and 1,145 Phc1-binding regions, to 791, 1,751, and 1,034 target genes, respectively (*SI Text* and Data Sets 8–10). The candidate Oct4-bound and Nanog-bound gene sets are enriched in Oct4-sorted+ genes ($P < 10^{-22}$, enrichment level $R > 2.3$, computations of P and R are given in *SI Text*), whereas Phc1-bound genes are depleted in Oct4-sorted+ genes ($P < 10^{-6}$, $R = 0.5$) and enriched in Oct4-sorted− genes ($P < 10^{-17}$, $R = 2.0$) (Fig. 2). In contrast, the overlaps between Oct4-sorted− genes and Oct4- or Nanog-bound genes are much less significant (P and R reported in Fig. 2). Similar intersection analysis of the Oct4-Ri± and RA± gene sets to the target genes of the TF-binding regions are presented in SI Fig. 5 (*A–F*). Together, these analyses suggest that Oct4 and Nanog are involved in the activation of many genes in ESC through direct binding to enhancers, and that, in a global sense, their involvement in gene repression may be less direct.

Only a minority (25%) of the Oct4-bound genes are identified by expression analysis as up-regulated in ESC (i.e., as Oct4+ genes). Conversely, most Oct4+ genes are not bound by Oct4. This phenomenon is expected. Oct4 occupancy on DNA may sometimes have no regulatory consequence for many reasons, including the lack of a necessary cofactor, or that competing regulatory mechanisms may be at work. However, the up-regulation of a gene in Oct4-high samples may be mediated by actions of factors other than Oct4. Although not surprising, the small (but significant) overlap between the bound genes and Oct4+ genes (SI Fig. 5 *G* and *H*) does suggest that neither expression data alone nor binding data alone are sufficient to identify cis-regulatory interactions, but rather, an integrative computational analysis is necessary to extract the full information. To support this analysis, we combine the expression and TF-binding data to obtain 219 Oct4-bound activator regions defined as Oct4-bound regions associated with Oct4+ genes, and 542 Nanog-bound activator regions defined similarly.

cis-Regulatory Analysis of Genes Expressed in ESC. To identify DNA motifs that may mediate cis-regulatory interactions in ESC, we searched TRANSFAC and surveyed the literature for motifs that are recognized by the TF in our Oct4-sorted+ set. In addition, although Stat3 and Sall4 are expressed but not up-regulated in ESC, we included them in our analysis. Stat3 is an established ESC regulator (2), and Sall4 has recently been found to have a role in

Table 1. Significance and enrichment levels of identified core regulators in mouse ESC

TF (fold change)	Oct4-bound activator region	Nanog-bound activator region	Phc1-bound region*	Conserved upstream region*
Oct4 (34)	1E-20,2.3/3.7	1E-8,1.5/1.8		6E-16,1.8
Sox2 (9.7)	2E-2,1.2/1.8	2E-12,1.5/2.3		2E-4,1.3
Nanog (52)		4E-3,1.3/1.7		1E-5,1.6
Stat3 (0.3)	1E-2,1.3/1.9	4E-4,1.3/1.0		2E-3,1.3
Esrrb (53)	8E-5,1.5/2.2	3E-6,1.4/1.4		
Sall4 (1.1)	6E-3,1.3/1.2			
Nr5a2 (12)	3E-3,1.3/1.6	4E-3,1.2/1.2		
Otx2 (3.9)	4E-3,1.4/1.7			2E-3,1.3
Tcf7 (4.6)		1E-3,1.3/1.4		
Etv5 (4.7)		2E-3,1.2/1.2		
Utf1 (20)				1E-4,1.4
Tcfap2c (23)				5E-3,1.3
Mtf2 (3.8)			3E-8,1.7	
Rest (4.5)			1E-3,1.5	
Rbpsuh (3.0)			3E-3,1.4	
Summary	88%	88%	57%	55%

The fold changes listed are those of the TF in the Oct4-sorted series. For the Oct4- and the Nanog-bound activator regions, the *P* values before conservation filtering, and the enrichment levels without/with conservation filtering, are reported. The Summary row reports the fraction of the regions with at least one scanned site of the indicated motifs.

*Regions are associated with the Phc1-repressed genes.

ESC (8). Because there are no known Sall4-binding sites, we inferred a provisional Sall4 motif by analyzing a reported Sall4-bound region in the Oct4 epiblast enhancer (ref. 14 and *SI Text*). Together these efforts resulted in 24 motifs (SI Table 2).

We use motif site enrichment analysis to see whether any of these motifs co-occur with the Oct4 motif in cis-regulatory modules. Specifically, we scanned Oct4-bound activator regions for sites of each motif in SI Table 2 and compared the count of detected sites to what is expected from a Poisson model estimated from control regions (*SI Text*). With a *P* value cutoff of 0.01, we found five motifs to be significantly enriched, namely Oct4 ($P = 1 \times 10^{-20}$, enrichment level $R = 2.3$, see *SI Text* of *P* and *R*), Esrrb ($P = 8 \times 10^{-5}$, $R = 1.5$), LRH-1/Nr5a2 ($P = 3 \times 10^{-3}$, $R = 1.3$), Otx2 ($P = 4 \times 10^{-3}$, $R = 1.4$), and Sall4 ($P = 6 \times 10^{-3}$, $R = 1.3$). Because the expected number of falsely detected motifs is <0.25 (24×0.01), the five detected motifs are not likely to be false positives. Esrrb is an orphan nuclear receptor whose expression starts ≈6.5 dpc. In the postgastrulation embryo, it is specifically expressed in primordial germ cells (15). In our Oct4-sorted series, it was highly expressed in Oct4-high cells and was down-regulated 53-fold in Oct4-low cells (i.e., fold change = +53). Finally, it was identified by an RNAi screen as an essential gene for self-renewal of ESC (3). LRH-1, with a fold change of +12, is another orphan nuclear receptor recently found to bind to the Oct4 epiblast enhancer and is required for the maintenance of Oct4 expression in the epiblast stage embryo (5). Sall4 is a zinc finger TF required for inner cell mass proliferation, and null mutants die soon after implantation (8). It has been found to co-occupy several cis-regulatory regions with Nanog in ESC (16). However, the Sall4 motif used in this analysis is computationally predicted; therefore, the above finding of Sall4 site enrichment is provisional in nature and needs to be tested carefully in future experiments. Although Oct4 is extremely highly enriched, the above list does not include the established ESC regulators Sox2 ($P = 0.019$, $R = 1.2$), Nanog ($P = 0.33$, $R = 1.1$), or Stat3 ($P = 0.014$, $R = 1.3$). Thus our initial cutoff may be too stringent. This motivated us to use cross-species conservation to determine whether TF with marginal *P* values (defined as those *P* values within twofold of the cutoff threshold, i.e., between 0.005 and 0.02) should be included in further analysis. We scanned for motif sites in only the part of each Oct4-bound activator region that is conserved, defined as the top 20% within that region in terms of the phastCons score (17) of University of California, Santa Cruz (UCSC) genome center (*SI Text*). For any motif with *P* value between 0.01 and 0.02 and whose enrichment in the conserved part increases by 20% or more relative to its overall enrichment level in Oct4-bound activator regions, we added it to the list of selected motifs for further analysis. Conversely, for any motif with *P* value between 0.005 and 0.01 and whose enrichment in conserved regions decreases by >20% relative to the overall level, we removed it from the selected motif list. After this conservation-filtering step, Sox2 (*R* increases from 1.2 to 1.8) and Stat3 (*R* increases from 1.3 to 1.9) were added to the selection to yield a final list of seven motifs that were significantly enriched in Oct4-bound activator regions. We note that the other four detected motifs with a *P* value < 0.005 (Oct4, Esrrb, LRH-1, and Otx2) all show higher levels of enrichment after conservation filtering (Table 1), although some of the *P* values after conservation filtering become larger because of the smaller counts involved (SI Table 3).

A similar analysis was performed on the Nanog-bound activator regions. At a *P* value cutoff of 0.01, we detected the motifs for Oct4 ($P = 1 \times 10^{-8}$, $R = 1.5$), Sox2 ($P = 2 \times 10^{-12}$, $R = 1.5$), Esrrb ($P = 3 \times 10^{-6}$, $R = 1.4$), Nanog ($P = 4 \times 10^{-3}$, $R = 1.3$), Etv5 ($P = 2 \times 10^{-3}$, $R = 1.2$), Tcf7 ($P = 1 \times 10^{-3}$, $R = 1.3$), Sall4 ($P = 6 \times 10^{-3}$, $R = 1.2$), Stat3 ($P = 4 \times 10^{-4}$, $R = 1.3$), and LRH-1 ($P = 4 \times 10^{-3}$, $R = 1.2$) as significantly overrepresented. However, when the top 20% conserved subregions were scanned, the Sall4-enrichment level decreased from 1.20 to 0.74. Because the original *P* value of Sall4 is only marginally significant (between 0.005 and 0.02), following the same criterion as in our analysis of the Oct4-bound activator regions, we removed Sall4, to obtain a final list of eight motifs that were significantly enriched in the Nanog-bound activator regions. It is comforting that the Nanog motif is enriched in these regions. Tcf7 is a transcriptional regulator downstream of Wnt signaling. Our result is thus consistent with the recent finding that activated Wnt signaling is sufficient for self-renewal in both human and mouse ESC (18). Interestingly, LRH-1, whose motif is again enriched, has been reported to work synergistically with the beta-catenin/T cell factor pathway to activate cyclin D1 and Myc in pancreatic and liver cell lines (19). Because beta-catenin is robustly expressed in all our ES and EB samples, it is possible that such a synergy is also at work in ESC.

Together, our analyses of the two sets of activator regions identified 10 TF that may function in gene activation in ESC. As expected, motifs for the master regulators Oct4, Sox2, and Nanog were identified, as was the motif for the well established regulator Stat3. In addition, we detected the motifs for Esrrb, LRH-1, Sall4, and Tcf7. Although recent publications have established the im-

portance of these four regulators in ESC, their functional roles remain to be investigated. Our result is consistent with the hypothesis that they may function as coactivators with the master regulators. The remaining two factors, Otx2 and Etv5, have not been implicated in ESC maintenance.

cis-Regulatory Analysis of Genes Repressed in ESC. Next we investigated the regulation of genes not expressed in ESC but activated upon differentiation. For this purpose, the Oct4- or Nanog-bound regions are not likely to be informative because they do not overlap significantly with Oct4-sorted−, RA+, or Oct4-Ri+ gene sets (Fig. 2 and SI Fig. 5). Instead, we rely on Phc1-bound regions that exhibit extremely significant overlap with Oct4-sorted− genes ($P = 9 \times 10^{-8}$, $R = 2.0$) and with RA+ genes ($P = 1 \times 10^{-33}$, $R = 3.1$) but not with Oct4-Ri+ genes (Fig. 2*C* and SI Fig. 5). We say that a gene is ESC-repressed if (*i*) it is in the Oct4-sorted− or RA+ gene sets and (*ii*) it has an average expression index below 150 in our three undifferentiated ESC samples. Motif site enrichment analysis (as described above) identified the motifs for Mtf2 ($P = 3 \times 10^{-8}$, $R = 1.7$), Rest ($P = 1 \times 10^{-3}$, $R = 1.5$), and Rbpsuh ($P = 3 \times 10^{-3}$, $R = 1.4$) as significantly enriched in Phc1-bound regions associated with ESC-repressed genes. Mtf2 is a zinc finger protein containing a PHD domain found in many chromatin modification factors. In view of its high expression in ESC and significant site enrichment on Phc1-bound regions, it is reasonable to hypothesize that Mtf2 may interact with polycomb complex in regulating histone modification. Rest, a kruppel-type zinc finger TF, plays a major role in the repression of neuronal genes in nonneuronal tissues and undifferentiated neural progenitors by recruiting Sin3/Hdac or CoRest repressor complexes (20). Both Sin3A and CoRest are up-regulated in Oct4-high cells, so these repressor complexes may also contribute to gene silencing in ESC. Rbpsuh is a DNA-binding regulator that activates Notch pathway target genes upon Notch signaling. However, in the absence of signaling, it serves as a transcriptional repressor through recruitment of the CtBP corepressor complex (21). In ESC the expression of all ligands and receptors in this pathway are very low, but both Rbpsuh and Ctbp2 are highly expressed in ESC and down-regulated but still expressed upon differentiation (SI Fig. 6). This is consistent with the hypothesis that Rbpsuh functions as a repressor in ESC and activator in RA-induced or EB differentiation.

To allow for the possibility that PRC1 may be recruited to the proximal promoter by regulators bound to more distal sites, we analyzed the conserved regions [top 20% in phastCons (17) score] within [−10 kb, +5 kb] around the transcription start sites of the Phc1-bound ESC-repressed genes. Surprisingly, a large number of motifs are detected at a *P* value cutoff of 0.01. In addition to motifs for Oct4 ($P = 5.6 \times 10^{-16}$, $R = 1.8$), Sox2 ($P = 1.6 \times 10^{-4}$, $R = 1.3$), Nanog ($P = 1.2 \times 10^{-5}$, $R = 1.6$), Stat3 ($P = 2 \times 10^{-3}$, $R = 1.3$), and Otx2 ($P = 2 \times 10^{-3}$, $R = 1.3$), we found that the motifs for Utf1 ($P = 1.2 \times 10^{-4}$, $R = 1.4$) and Tcfap2c ($P = 5 \times 10^{-3}$, $R = 1.3$) are also enriched. Utf1 is a well established ESC marker with no known ESC functions. Tcfap2c is essential in the extraembryonic lineages in early postimplantation development (22) and has also been found to have a role in the maintenance of a proliferative and undifferentiated state of cells (23). How these seven motifs regulate Phc1-bound ESC-repressed genes may be complex. A cis-regulatory module involving these motifs may facilitate recruitment of polycomb complexes by ESC regulators, or it may mediate rapid initiation of gene transcription by ESC regulators still present when polycomb repression is terminated in some early differentiated lineages. Consistent with both of these scenarios was our finding that Phc1-bound genes overlap significantly with Oct4-bound or Nanog-bound genes ($P = 5 \times 10^{-6}$ and 6×10^{-4}, respectively), showing that in ESC, Oct4 and Nanog tend to occupy enhancers associated with genes bound by PRC1. Finally, it is also possible that some of the motifs (e.g., the Stat motif) may be used by TF not in the Oct4+ set, to activate transcription during ESC differentiation (e.g., other Stat proteins up-regulated upon differentiation). In view of the critical developmental roles of the target genes of PRC1 repression, it will be important to resolve these different possibilities by further experimental studies.

An Oct4–Sox2–Nanog Regulatory Network in Mouse ESC. In total, the above analyses identified 15 regulators (Table 1 and SI Fig. 7), called core regulators hereafter, which may be involved in the maintenance of ESC. To identify the regulatory interactions among them, we first analyzed the Oct4-bound regions associated with each of these 15 genes to identify all high-quality Oct4 sites and Oct4–Sox2 double sites *de novo* discovered (*SI Text*) on these regions. Regarding these sites as anchors, we further identified sites of other core regulators that lie within 150 bp of the anchor sites. This window size was chosen by our prior experience that the expected length of a cis-regulatory module is ≈300 bp. In Fig. 3, a blue arrow from Oct4/Sox2 to a target gene indicates an inferred regulatory interaction between Oct4/Sox2 and the target gene, as supported by the existence of an Oct4/Oct4–Sox2 site on an Oct4-bound activator region associated with the target gene. Similarly, a pink arrow indicates a potential regulatory interaction between an Oct4-coregulator (Esrrb, LRH-1/Nr5a2, etc.) and a target gene. In a similar manner, we also analyzed the Nanog-bound regions associated with these 15 genes, where Nanog, Sox2, and Oct4 sites were regarded as anchor sites. The reason is that Sox2 ($P = 2 \times 10^{-12}$) and Oct4 ($P = 1 \times 10^{-8}$) are by far the most significant motifs found on the Nanog-bound activator regions, whereas the Nanog motif itself is only moderately enriched ($P = 4 \times 10^{-3}$). The directed graph of these 15 regulators and Phc1 in Fig. 3 (the pink circles and the directed edges connecting them) can be viewed as a representation of the core gene regulatory network in ESC centering on the three master regulators Oct4, Nanog, and Sox2. We note that two core regulators, Utf1 and Mtf2, identified as repressive TF on Phc1-bound regions, are not included in the core network, because we did not detect any anchor sites in the Oct4- or Nanog-bound regions associated with them. Utf1 has only one associated Nanog-bound region at downstream 6 kb of its transcription start sites, which does not cover the Oct4–Sox2 double site reported previously (24). Mtf2 is associated with one Oct4-bound and one Nanog-bound region, but both of them are almost purely repeats. In addition to the core regulators, the network can be expanded to include all 334 putative target genes (of the network) defined as those genes associated with Oct4- or Nanog-bound activator regions that contain at least one anchor site. Adding these putative targets resulted in an expanded network of 337 genes after elimination of redundancy (SI Table 4).

Three lines of independent experimental evidence were used to assess the relevance of this predicted network in an unbiased way. First, among the six predicted regulatory interactions between the three master regulators Oct4, Nanog, and Sox2, five of them were supported by reported experimental validations (ref. 25 and its relevant references). Secondly, our network predicted 242 and 96 target genes of Sox2 and Nanog, respectively (SI Table 4), among which 59 and 31 were identified in ref. 3 as genes with expression affected by Sox2 and Nanog RNAi knockdown experiments. These results correspond to very significant overlaps and high enrichment ratios ($P < 6 \times 10^{-14}$, $R > 3.9$). Note that these expression profiles were not included in our analysis and thus can be used as independent validation data. Finally, we turn to a recent study of protein interaction network in ESC (26). By using affinity purification of Nanog or Oct4 followed by mass spectrometry, this study identified 23 high-quality protein-interaction partners of Nanog or Oct4 (figure 4*b* in ref. 26), and 8 of them turned out to be in our expanded network: Sall4, Esrrb, Rest, Rybp, Zfp219, Dax1/Nr0b1, Sall1, and Rif1. Thus, although our network is predicted only on the basis of expression and location data, it captures ≈34% of the high-quality protein interaction partners of Nanog or Oct4. This result represents an extremely significant enrichment ($P = 6 \times 10^{-9}$), especially

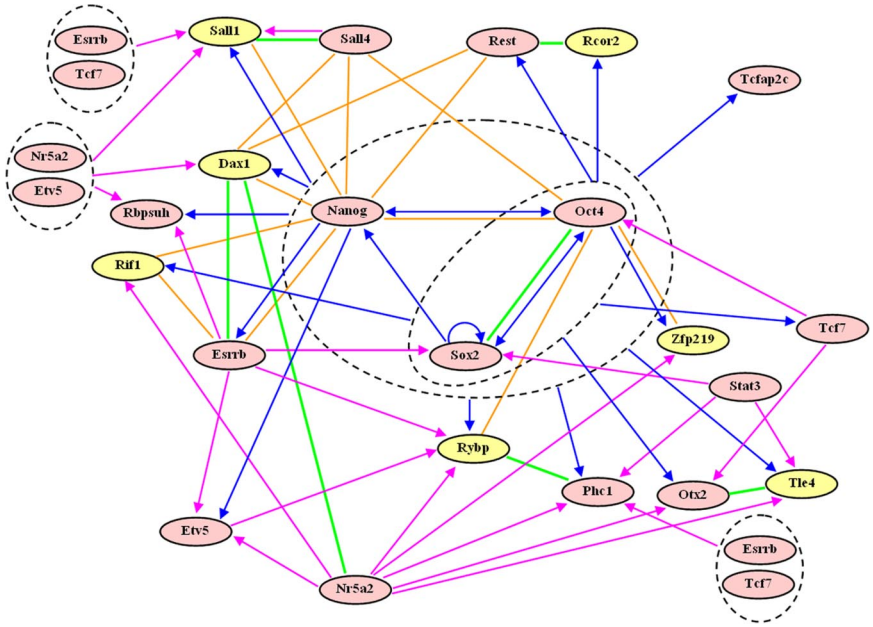

Fig. 3. A regulatory network in mouse ESC anchored on the master regulators Oct4, Sox2, and Nanog. The network represents the interactions among the core regulators (pink) and their protein-interaction partners (yellow). Blue and pink arrows indicate regulatory interactions inferred by anchor sites and by sites of coregulators within 150 bp of the anchor sites, respectively. Orange and green lines represent protein interactions identified in ref. 26 and reported in the literature, respectively. Arrows from a dashed ellipse indicate that the targets are regulated by all of the regulators inside the ellipse. Some regulators appear multiple times in the network to reduce the number of intersecting arrows.

so given that a protein-interaction partner of Nanog or Oct4 is not necessarily a transcriptional target of these regulators. We add these interaction partners to our network of core regulators. In addition, Tle4 and Rcor2 are also added because of their known interactions with the core regulators Otx2 and Rest, respectively (20, 27). The resulting network of core regulators and their protein-interaction partners are shown in Fig. 3.

As expected, at the heart of the network are the three master TF that regulate each other and other genes in the network. Joining them to drive the regulatory network is a group of coregulators, Esrrb, Tcf7, Sall4, LRH-1/Nr5a2, Stat3, and Etv5. Esrrb and Tcf7 are particularly noteworthy because they not only are direct targets of the master regulators but also participate in the regulation of them. Finally, there are many genes at the receiving ends of the regulatory interactions originating from Oct4, Sox2, Nanog, and their coregulators. Of these, Phc1, Rybp, Rbpsuh, Sall1, Tle4, and Dax1 have a large number of incoming arrows, suggesting that the robust regulation of these genes may be essential for ESC maintenance and pluripotency. Phc1 is the target of seven core regulators. Of the many polycomb complex PRC1 components, Phc1 is the most strongly regulated at the transcriptional level with the highest expression fold change in our Oct4-sorted profiles (SI Data Sets 1 and 2). Furthermore, Rybp, a known protein-interaction partner of multiple components of the PRC1 complex that contains Phc1, is also strongly regulated by Oct4, Sox2, Nanog, and three other core regulators. The Notch-responsive regulator Rbpsuh is the target of six core regulators. As discussed above, Rbpsuh may modulate polycomb regulation of developmental genes by serving as a transcriptional switch dependent on Notch signaling. Given the importance of the developmental genes silenced by PRC1, it is not surprising that multiple and perhaps redundant pathways are used to regulate their expression. Sall1, an interaction partner of Nanog and a known coactivator with Sall4 (8), is the target of eight core regulators. In contrast, although Sall4 is a core regulator and physically interacts with both Nanog and Oct4, it does not seem to be strongly regulated by the network. It would be interesting to investigate whether the regulatory network may regulate Sall4 indirectly through Sall1. The Gruocho-like corepressor Tle4 is the target of five core regulators, and it interacts with the core regulator Otx2 (24). Otx2, itself a target of four core regulators, is a bicoid-class TF essential in primitive streak organization and anterior neural development (28). In both the RNAi and RA series, Otx2 has high fold change in favor of ESC. In the Oct4-sorted series, Otx2 expression level in 2- and 4-day EB is significantly up- or down-regulated compared with ESC, depending on whether it is Oct4-high or Oct4-low. Thus the role of Otx2 may be to work with Oct4 to maintain gene expression in early progenitors of ectodermal lineages. Finally, the orphan nuclear receptor Dax1/Nr0b1 is the target of five core regulators, and it also interacts at the protein level with Nanog, Sall4, Rest, Esrrb, and Nr5a2. Furthermore, Dax1 has an impressive fold change of 51 in favor of Oct4-high cells in our Oct4-sorted series. Thus it is likely a critical component of the regulatory network. An important finding of our study is the surprisingly prominent roles of nuclear receptors in the network. Both Esrrb/Nr3b2 and LRH-1/Nr5a2 participate in the control of many core regulators, and Esrrb has the most significant site enrichment ($P = 10^{-5}$ to 10^{-6}) in the Oct4- and Nanog-bound activator regions after Oct4 and Sox2. By its interaction with Esrrb and Nr5a2, Dax1 may facilitate the cooperation of these important core regulators with the master regulator Nanog. However, it has been shown that Dax1 may inhibit the transactivator activities of its nuclear receptor partner (29), which is contrary to the putative activator roles of Esrrb and LRH-1. The resolution of this issue through experimental investigation may hold the key to the full understanding of nuclear receptor functions in the regulatory network.

We now turn to the remaining part of the expanded network that includes all putative target genes, for which we present the regulatory interactions in SI Table 4. The high ranking target genes with $P < 0.001$ (*SI Text*) are presented in SI Table 5. It is remarkable that 45% of the genes in this table are transcriptional regulators. This result suggests that the most important direct targets activated by the core regulatory network are themselves transcriptional regulators whose activities may extend the regulatory effects of the

network to numerous secondary targets. Some of these high ranking targets, namely Zfp219, Rif1, Phc1, Rbpsuh, Tle4, Sall1, Rybp, and Otx2, are themselves part of the core network in Fig. 3, whereas others, such as Klf4, Lefty1, Myc, Trp-53, and Foxd3, have previously been implicated to have roles in ESC (4, 6, 7). Given the high enrichment of genes known to be important to ESC, it is likely that many of the remaining genes in this table may also have functional roles in ESC. To give one example, the high-ranking gene Jarid2 is the target of seven core regulators. Jarid2, also known as Jumonji, is highly expressed in ESC and down-regulated ≈10-fold in Oct4-low EB cells. Its domain structure suggests that it is a chromatin-associated factor with histone modification activities (30). As such, Jarid2 may provide yet another means by which the core network regulates chromatin status in ESC. Full lists of target genes are provided in SI Tables 6–9.

Discussion

Combining information from global expression and location data, we have inferred a gene regulatory network in ESC composed of Oct4, Nanog, Sox2, and their coregulators. Our analysis provided overwhelming evidence of site enrichment for Oct4 in both Oct4- and Nanog-bound activator regions, and for Sox2 in Nanog-bound activator regions. In contrast, Nanog sites are only moderately enriched even in Nanog-bound regions. The lack of site enrichment for Nanog may be due to an inaccurate weight matrix or due to weak DNA-binding activity of Nanog. We favor the second explanation because, even with extensive refinement of the weight matrix, the enrichment level of Nanog sites is still unimpressive. Our analysis also revealed extremely significant site enrichment of Oct4 and its coregulators in the conserved upstream regions of Phc1-bound ESC-repressed genes, raising the possibility that cis-elements containing sites for these regulators may have a role in recruiting or modulating polycomb function. In addition, site enrichment analysis of Phc1-bound promoters of ESC-repressed genes suggested Mtf2, Rbpsuh, and Rest as potential corepressors on these promoters. On the basis of detailed mapping of binding sites of master regulators and their coregulators, we inferred a tentative gene regulatory network in ESC. Although preliminary and incomplete, this network has already revealed extensive cross-regulatory interactions among the core regulators. It also identified many downstream regulators that propagate the effects of the core network. Several of the most regulated targets by this network, such as Phc1, Rybp, and Jarid2, are involved in the epigenetic control machinery. Our analysis also revealed the surprisingly prominent roles of several orphan nuclear receptors in the network. To our knowledge, this is the first systematic global inference of regulatory interactions in ESC based on integrated analysis of expression, location, and genome sequence data. With this approach, it should be possible to expand, revise, and refine the network as more expression and location data become available in the near future.

Materials and Methods

Oct4-GFP mES (11) were cultured on mouse embryonic fibroblasts according to standard methods. Free-floating EB were formed by passaging the cells once onto gelatin-coated plates to deplete the feeders, and when they became semiconfluent, they were passaged into 6-well ultra-low cluster plates (Costar, Cambridge, MA). Differentiation medium was DMEM with 15% FCS, 1× Glutamax, and 1× nonessential amino acids (Gibco, Carlsbad, CA). Cystic EB became apparent after 4–6 days, with extensive differentiation as evidenced by areas of beating cardiac myocytes after 6–8 days. At various time points, EB were dissociated to single cells by using trypsin/EDTA and sorted into GFP± fractions by using a MoFlo cell sorter (Dako North America, Carpinteria, CA). The sorting gates were set by using parallel EB from a non-GFP mES line as a negative control, and for each sample 0.5–1.0 × 10^6 cells were recovered. Replicates represent independent differentiations performed at different times. RNA was isolated from the sorted cells by using an RNeasy kit (Qiagen, Valencia, CA). RNA samples were prepared for hybridization by using One Cycle target labeling reagents (Affymetrix, Santa Clara, CA), hybridized to Mouse Genome 430 V.2 Affymetrix microarrays and scanned by using a GeneChip scanner 3000 7G all according to the manufacturer's protocols.

Details of the computational methods for data analysis are given in *SI Text*.

We thank Hong Yang, Hongkai Ji, Wenxiu Ma, and Sheng Zhong for helpful discussions. This work was supported by National Institutes of Health Grant GM067250.

1. Chambers I, Smith A (2004) *Oncogene* 23:7150–7160.
2. Niwa H, Burdon T, Chambers I, Smith A (1998) *Genes Dev* 12:2048–2060.
3. Ivanova N, Dobrin R, Lu R, Kotenko I, Levorse J, DeCoste C, Schafer X, Lun Y, Lemischka I (2006) *Nature* 442:533–538.
4. Hanna LA, Foreman RK, Tarasenko IA, Kessler DS, Labosky PA (2002) *Genes Dev* 16:2650–2661.
5. Gu P, Goodwin B, Chung AC, Xu X, Wheeler DA, Price RR, Galardi C, Peng L, Latour AM, Koller BH, *et al.* (2005) *Mol Cell Biol* 25:3492–3505.
6. Takahashi K, Yamanaka S (2006) *Cell* 126:1–14.
7. Lin T, Chao C, Saito S, Mazur SJ, Murphy ME, Appella E, Xu Y (2005) *Nat Cell Biol* 7:165–171.
8. Sakaki-Yumoto M, Kobayashi C, Sato A, Fujimura S, Matsumoto Y, Takasato M, Kodama T, Aburatani H, Asashima M, Yoshida N, *et al.* (2006) *Development (Cambridge, UK)* 133:3005–3013.
9. Nichols J, Zevnik B, Konstantinos A, Niwa H, Klewe-Nebenius D, Chambers I, Scholer H, Smith A (1998) *Cell* 95:379–391.
10. Niwa H, Miyazaki J, Smith A (2000) *Nat Genet* 24:372–376.
11. Yeom YI, Fuhrmann G, Ovitt CE, Brehm A, Ohbo K, Gross M, Hubner K, Scholer HR (1996) *Development (Cambridge, UK)* 122:881–894.
12. Loh YH, Wu Q, Chew JL, Vega VB, Zhang W, Chen X, Bourque G, George J, Leong B, Liu J, *et al.* (2006) *Nat Genet* 38:431–440.
13. Boyer LA, Plath K, Zeitlinger J, Brambrink T, Medeiros LA, Lee TI, Levine SS, Wernig M, Tajonar A, Ray MK, *et al.* (2006) *Nature* 441:349–353.
14. Zhang J, Tam WL, Tong GQ, Wu Q, Chan HY, Soh BS, Lou Y, Yang J, Ma Y, Chai L, *et al.* (2006) *Nat Cell Biol* 8:1114–1123.
15. Mitsunaga K, Araki K, Mizusaki H, Morohashi K, Haruna K, Nakagata N, Giguere V, Yamamura K, Abe K (2004) *Mech Dev* 121:237–246.
16. Wu Q, Chen X, Zhang J, Loh YH, Low TK, Zhang W, Zhang W, Sze SK, Lim B, Ng HH (2006) *J Biochem* 281:24090–24094.
17. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, *et al.* (2005) *Genome Res* 15:1034–1050.
18. Sato N, Meijer L, Skaltsounis L, Greengard P, Brivanlou AH (2005) *Nat Med* 10:55–63.
19. Bortrugno OA, Fayard E, Annicotte JS, Haby C, Brennan T, Wendling O, Tanaka T, Kodama Y, Thomas W, Auwerx J, *et al.* (2004) *Mol Cell* 15:499–509.
20. Lunyak V, Burgess R, Prefontaine GG, Nelson C, Sze SH, Chenoweth J, Schwartz P, Pevzner PA, Glass C, Mandel G, *et al.* (2002) *Science* 298:1747–1752.
21. Oswald F, Winkler M, Cao Y, Astrahantseff K, Bourteele S, Knochel W, Borggrefe T (2005) *Mol Cell Biol* 25:10379–10390.
22. Auman HJ, Nottoli T, Lakiza O, Winger Q, Donaldson S, Williams T (2002) *Development (Cambridge, UK)* 129:2733–2747.
23. Jager R, Werling U, Rimpf S, Jacob A, Schorle H (2003) *Mol Cancer Res* 1:921–929.
24. Nishimoto M, Fukushima A, Okuda A, Muramatsu M (1999) *Mol Cell Biol* 19:5453–5465.
25. Chew JL, Loh YH, Zhang W, Chen X, Tam WL, Yeap LS, Li P, Ang YS, Lim B, Robson P, *et al.* (2005) *Mol Cell Biol* 14:6031–6046.
26. Wang J, Rao S, Chu J, Shen X, Levasseur DN, Theunissen TW, Orkin SH (2006) *Nature* 444:364–368.
27. Puelles E, Annino A, Tuorto F, Usiello A, Acampora D, Czerny T, Brodsky C, Ang SL, Wurst W, Simeone A (2004) *Development (Cambridge, UK)* 131:2037–2048.
28. Ang SL, Jin O, Rhinn M, Daigle N, Stevenson L, Rossant J (1996) *Development (Cambridge, UK)* 122:243–252.
29. Suzuki T, Kasahara M, Yoshioka H, Morohashi K, Umesono K (2003) *Mol Cell Biol* 23:238–249.
30. Tsukada Y, Fang J, Erdjument-Bromage H, Warren ME, Borchers CH, Tempst P, Zhang Y (2006) *Nature* 439:811–816.