BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 17 2013, pages 2162–2168
doi:10.1093/bioinformatics/btt365

Gene expression

Advance Access publication June 21, 2013

Co-regulation in embryonic stem cells via context-dependent binding of transcription factors

Yuju Lee and Qing Zhou*

Department of Statistics, University of California, Los Angeles, CA 90095, USA

Associate Editor: Inanc Birol

ABSTRACT

Motivation: With the accumulation of genome-wide binding data for many transcription factors (TFs) in the same cell type or cellular condition, it is of great current interest to systematically infer the complex regulatory logic among multiple TFs. In particular, ChIP-Seq data have been generated for 14 core TFs critical to the maintenance and reprogramming of mouse embryonic stem cells (ESCs). This provides a great opportunity to study the regulatory collaboration and interaction among these TFs and with other unknown co-regulators.

Results: In combination with liquid association among gene expression profiles, we develop a computational method to predict context-dependent (CD) co-egulators of these core TFs in ESCs from pairwise binding datasets. That is, co-occupancy between a core TF and a predicted co-regulator depends on the presence or absence of binding sites of another core TF, which is regarded as a binding context. Unbiased external validation confirms that the predicted CD binding of a co-regulator is reliable. Our results reveal a detailed CD co-regulation network among the 14 core TFs and provide many other potential co-regulators showing strong agreement with the literature.

Availability: See www.stat.ucla.edu/~zhou/CMF for software and source code.

Contact: zhou@stat.ucla.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 5, 2013; revised on June 5, 2013; accepted on June 19, 2013

1 INTRODUCTION

Many efforts have been made recently to identify the core regulatory network in mouse embryonic stem cells (ESCs; Chen *et al.*, 2008; Ivanova *et al.*, 2006; Kim *et al.*, 2008; Zhou *et al.*, 2007). The network is anchored on three master regulators, namely, Oct4 (Pou5f1), Sox2 and Nanog. On the other hand, four transcription factors (TFs), Oct4, Sox2, cMyc and Klf4, can reprogram somatic cells back to ESC-like cells having the characteristics of self-renewal and pluripotency (Takahashi and Yamanaka, 2006). Genome-wide ChIP-Seq data provide the binding regions of >10 TFs that play key regulatory roles in mouse ESCs (Chen *et al.*, 2008; Heng *et al.*, 2010; Marson *et al.*, 2008). Computational analyses have been performed to identify the sequence motifs of these core TFs (Bailey, 2011; Mason *et al.*, 2010; Thomas-Chollier *et al.*, 2012) and to detect other co-regulators that may regulate genes by working with these TFs (Chen and Zhou, 2011; He *et al.*, 2009). A common observation on these binding data is the partial and significant overlapping between the binding regions of almost any pair of the TFs, suggesting extensive co-regulatory interactions among all of them. In general, one may define three sets of binding regions given two TFs, X and Y, that is, the co-bound regions S_{XY}, and the regions bound only by X or Y, denoted by $S_{X\setminus Y}$ and $S_{Y\setminus X}$, respectively. From the perspective of TF X, the regions S_{XY} and $S_{X\setminus Y}$ can be regarded as two different binding contexts, defined by whether binding is accompanied by TF Y.

In our previous work (Mason *et al.*, 2010), we developed a specific motif discovery method, the contrast motif finder (CMF), which finds *de novo* motifs by contrasting two sets of sequences, say S_{XY} and $S_{X\setminus Y}$. Motifs found by CMF are differentially enriched between the two sequence sets. Applying CMF to contrast co-bound sequences of Oct4 and Sox2 against sequences bound only by Oct4, we discovered novel context-dependent (CD) motifs for Oct4, that is, the motif pattern recognized by Oct4 depends on the nearby Sox2 binding (Mason *et al.*, 2010). In this work, we generalize this idea to the prediction of CD co-regulators of a TF. For a TF X with binding context defined by another TF Y, we aim at finding motifs that show differential enrichment between S_{XY} and $S_{X\setminus Y}$. These motifs may be bound by co-regulators of TF X, dependent on the binding of Y. If a co-regulator Z only works with X when Y is present, one may expect to find the motif of Z in S_{XY} but not in $S_{X\setminus Y}$. On the contrary, if Z collaborates with X when Y is absent, then the motif of Z will be enriched in $S_{X\setminus Y}$ but not in S_{XY}. See Figure 1a for an illustration. The motif of such a CD co-regulator can be reliably found by CMF when contrasting the two sets of sequences. On the other hand, gene expression profiles of the three TFs, X, Y and Z, may provide independent evidence for the potential CD co-regulation. If Z is a co-regulator of X when Y is present (or absent), one expects to observe strong positive correlation between the expression profiles of X and Z when the expression level of Y is high (or low) (Fig. 1b). Such a three-way correlation pattern can be measured by liquid association (LA) developed by Li (2002).

By combining CD binding and LA among three TFs, we restrict our analysis to some specific CD co-regulation patterns. We assume that binding site patterns are identical across all samples from which gene expression data are collected. Use the right panels in Figure 1a and b to demonstrate our model. The binding pattern that X co-binds, respectively, with Z and Y for different target genes is assumed to be true across all samples. Together with the three-way expression pattern among the three TFs, we are considering the following co-regulatory relationship. When

*To whom correspondence should be addressed.

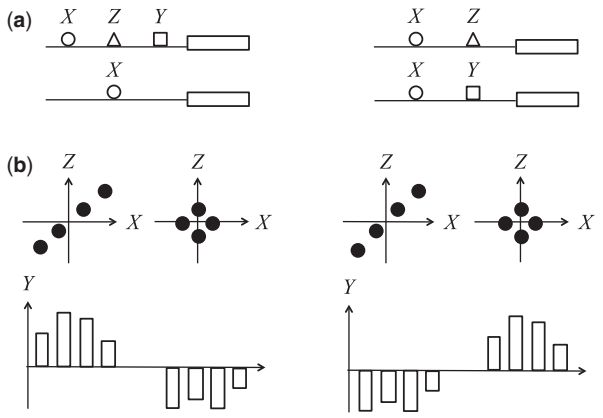

Fig. 1. CD co-regulation. The TF Z is a CD co-regulator of X with the context defined by the binding of Y. (**a**) The left (right) panel illustrates the scenario in which Z co-binds with X when Y binding is present (absent). (**b**) Correspondingly, the correlation between the expression profiles of X and Z is strong when Y has high (left panel) or low (right panel) expression. The expression levels of X and Z are plotted as solid circles and those of Y are represented by vertical bars. Note that expression profiles have been centered to have mean zero

the expression of Y is low but that between X and Z is highly correlated, the co-occurrence between the sites of X and Y is not functional (because of a lack of Y expression), while the co-binding between X and Z regulates their target genes. When high expression of Y with low correlation between the expression of X and Z is observed, the opposite scenario is considered. This is of course a simplified model, which does not include other types of more general conditional co-regulation. But for the data we analyzed in this work, a substantial fraction of TF triplets with CD binding in their ChIP-Seq data indeed show significant LA among their expression profiles. See Section 2.3 for details.

2 RESULTS

We performed a systematic study on 14 TFs (Table 1) with available ChIP-Seq data (Chen *et al.*, 2008; Heng *et al.*, 2010; Marson *et al.*, 2008) to find CD co-regulators in mouse ESCs. All of the 14 TFs play critical or relevant regulatory roles in ESCs. For instance, Oct4, Sox2, Nanog, Esrrb and Zfx are known regulators of pluripotency and/or self-renewal (Chen *et al.*, 2008); LIF signal pathway can sustain self-renewal of cells by activating Stat3 (Niwa *et al.*, 1998); together with Oct4 and Sox2, cMyc and Klf4 can reprogram somatic cells to pluripotent cells (Takahashi and Yamanaka, 2006). We call them core TFs hereafter. Expression data of Ivanova *et al.* (2006) and Zhou *et al.* (2007) were integrated for LA analysis. Ivanova *et al.* (2006) reported expression data on 70 samples in response to RNAi of a few key TFs and during retinoic acid induction of mouse ESCs. Zhou *et al.* (2007) generated 16 expression profiles, including 3 profiles of undifferentiated mouse ESCs, 5 profiles from early embryoid body (EB) with high Oct4 expression, and 8 profiles from EB with low Oct4 expression.

To unify description and simplify notations, we present our results on the 182 ($= 14 \times 13$) ordered pairs among the 14 core TFs. For an ordered pair (X, Y), our goal is to predict co-regulators Z of X with context defined by the presence or absence of Y binding sites. Two sets of sequences, S_{XY} and $S_{X\setminus Y}$, were extracted from the ChIP-Seq data of X and Y for each ordered pair (more details in 'Methods' section). See Figure 2 for a flowchart including the key steps of our analysis.

2.1 Contrast motif finding

We input S_{XY} and $S_{X\setminus Y}$ to CMF if both sequence sets have at least 300 sequences, as *de novo* motif finding with a large number of sequences is often more reliable. This requirement cut down the number of ordered pairs to 172. We ran CMF to find at most 60 motifs for each pair of input sequence sets. For a found motif, CMF outputs a list of predicted binding sites in each of the two input sequence sets and computes a likelihood ratio (LR) score for each predicted site. Usually, a motif found by CMF is only enriched in one of the two sequence sets and the predicted sites in the other set may serve as a control set. We performed a two-sample t-test on the log LR scores from the two sets to determine the significance level of differential enrichment. Consequently, we obtained a t-statistic with a P-value for each motif found by CMF. The t-statistic is >0 (<0) if the motif is more enriched in S_{XY} ($S_{X\setminus Y}$).

When contrasting between S_{XY} and $S_{X\setminus Y}$, CMF is expected to find the motif of Y more enriched in the co-bound sequence set S_{XY}. Consensus motifs of the 14 core TFs (Table 1) were extracted from TRANSFAC (Matys *et al.*, 2003) and other literature. For 64 ordered pairs, CMF found the expected motif with at most one mismatch to the known consensus and with a significant t-statistic ($P<10^{-4}$). Because our goal is to identify motifs of co-regulators, we excluded from further analysis all motifs that match the consensus motifs, or other similar variants in TRANSFAC, of the 14 core TFs. To reduce false discoveries, we focused on motifs whose t-statistics were more significant than the t-statistic of at least one expected motif found from the same pair. Note that CMF may find multiple expected motifs for an ordered pair, all matching the consensus motif of Y.

Next, two criteria were applied to filter the motifs. First we filtered out a motif if its consensus sequence contains >80% of A/T or G/C. This removed AT-rich or GC-rich motifs, which are likely repetitive patterns, and left 372 unique motifs. As we are interested in CD regulation, we used a P-value cutoff of 1×10^{-4} for the t-statistic to filter out motifs that did not show significant enrichment difference between S_{XY} and $S_{X\setminus Y}$. It turned out that 339 motifs passed the P-value cutoff, which gives an extremely low false discovery rate (FDR) of 0.011% ($=372 \times 10^{-4}/339$). Hereafter, we call them post-filtering (PF) motifs. We classify all PF motifs into two categories: Y-present motifs, which are more enriched in S_{XY}, and Y-absent motifs more enriched in $S_{X\setminus Y}$.

2.2 Identifying CD co-regulators

We mapped PF motifs to known position-specific weight matrices (PWMs) in TRANSFAC (details in 'Methods' section), and according to its annotation 309 of the PF motifs may be recognized by at least one mouse TF. Given an ordered pair (X, Y), suppose we have found a Y-present motif that may be bound by a TF Z. Then our hypothesis is that the TF Z may co-regulate

Table 1. The 14 core TFs and their consensus motifs

Oct4 ATGCAAAT	Sox2 CTTTGTT	Nanog CGATTAAG	Esrrb AAGGTCA	Stat3 TTCCCGGAA
cMyc CACGTG	nMyc CACGTG	E2f1 CGCGCCAAA	Smad1 CAGACA	Zfx AGGCCT
Tcfcp2l1 AACCAGT	Klf4 GGGCGGG	Tcf3 TCAAAG	Nr5a2 CAAGGCC	

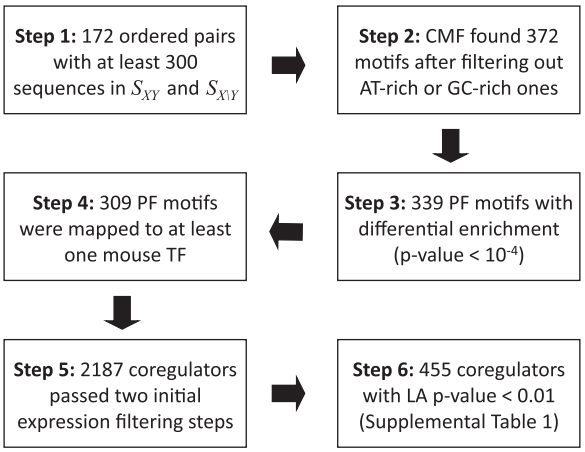

Fig. 2. Flowchart of six key steps for the prediction of CD co-regulators

genes with X when Y also binds the regulatory regions (left panel of Fig. 1a). In this scenario, we call Z a CD co-regulator of X when Y is present. Analogously, one may define a CD co-regulator of X when Y is absent (right panel of Fig. 1a).

Expression profiles may provide independent information on the hypothesized CD co-regulation. Denote by $\mathbf{x}, \mathbf{y}, \mathbf{z}$ the gene expression profiles across n conditions of the TFs, X, Y, Z, respectively. Assume that all the expression profiles have been standardized to have mean zero and a unit variance, and particularly, $\mathbf{y}$ has been transformed to follow the standard normal distribution. Then the LA between X and Z with respect to Y is measured by

$$LA(X, Z|Y) = \frac{1}{n}\sum_{i=1}^{n} x_i z_i y_i, \quad (1)$$

where x_i is the i^{th} component of $\mathbf{x}$ and similarly for y_i and z_i. Intuitively, if $\mathbf{x}$ and $\mathbf{z}$ are highly correlated when y_i is positive, the LA will be large in magnitude. If the correlation between $\mathbf{x}$ and $\mathbf{z}$ is independent of $\mathbf{y}$, then $LA(X, Z|Y)$ will be close to zero. Therefore, if Z is a CD co-regulator of X when Y is present (absent), we expect $LA(X, Z|Y) > 0\,(<0)$. Moreover, for both types of CD co-regulation, the overall correlation between $\mathbf{x}$ and $\mathbf{z}$ should be positive. See Figure 1b for demonstration. A brief review of LA can be found in 'Methods' section.

Consequently, we discarded a candidate co-regulator Z if it showed negative expression correlation with X or if the sign of $LA(X, Z|Y)$ was inconsistent with the category of the motif (Y-present or Y-absent) in both expression datasets (Ivanova *et al.*, 2006; Zhou *et al.*, 2007). A total of 2187 candidate co-regulators, after removing redundant ones predicted in the same ordered pair, passed the above two filtering criteria. Then we calculated the P-value of the LA score (1) for each candidate co-regulator Z by permuting the expression profile of Y (Li, 2002). Using a p-value cutoff of 0.01, we obtained 455 co-regulators, over 48 ordered pairs, showing significant LA scores, with an expected FDR of 4.8% ($= 2187 \times 0.01/455$). For the full list of the predicted CD co-regulators, see Supplemental Table 1.

2.3 Validation by ChIP-Seq data

When we predicted a co-regulator Z for an ordered pair X and Y, we only used the ChIP-Seq data of X and Y and the gene expression profiles of the three TFs. Therefore, if available, ChIP-Seq data of Z can be used to validate our results. If Z is predicted as a co-regulator of X when Y is present (absent), then the ChIP-Seq peaks of Z should occur more frequently in S_{XY} ($S_{X\setminus Y}$). There are four cases (Table 2) in which the predicted co-regulators Z are among the 14 core TFs (and thus have ChIP-Seq data available) and their motifs (found by CMF) have at most one mismatch to the corresponding consensus sequences in Table 1. For the four cases, we calculated the occurrence rates per base pair of the ChIP-Seq peaks of Z in S_{XY} and in $S_{X\setminus Y}$. Then a P-value was determined by comparing the two rates according to the two-sample proportion test. We report in Table 2 the P-value together with the ratio of the peak occurrence rate in S_{XY} to that in $S_{X\setminus Y}$ for each case.

It is comforting to see that for all the four cases, the predicted CD co-regulation is validated by the peak occurrences with a significant P-value. Consistent with its predicted role as a Y-present co-regulator, nMyc shows much higher binding enrichment in the corresponding S_{XY} than in $S_{X\setminus Y}$ (here, X is Sox2 and Y is Zfx). The ratio of the occurrence rates of the ChIP-Seq peaks of this TF between the two sets of binding regions is more than five with extremely strong statistical significance. Similarly, one sees from the table a few predicted Y-absent co-regulators with strong support from their ChIP-Seq data. Overall, the high validation rate confirms the accuracy of our prediction of CD co-regulators.

Many of the 14 core TFs have multiple PWMs in TRANSFAC, although some of the matrices may have a lower quality. We found 44 predicted co-regulators whose motifs can match to at least one TRANSFAC PWM of the 14 TFs and performed the same ChIP-Seq data validation as for the above four cases. It turned out that 27 (61%) of the 44 predictions can be validated by the corresponding ChIP-Seq data (Supplemental Table 2). This again demonstrates the correctness of the majority of our predicted CD binding. Figure 3 shows the respective histograms of the ratio and the inverse ratio of the ChIP peak occurrence rates for the 23 Y-present and the 21 Y-absent co-regulators. We see that the validation rate is particularly high ($83\% = 19/23$) for Y-present co-regulators. Among the 19 co-regulators with this ratio >1, the P-value of the difference between peak occurrence rates is $< 2 \times 10^{-6}$ for 18 of them and is $\sim 7 \times 10^{-3}$ for the other one (Supplemental Table S2).

Using the ChIP-Seq data of the 14 core TFs, we examined the use of LA among gene expression in predicting CD co-regulators. We identified 45 triplets (X, Y, Z) for which the ratio of the

Table 2. ChIP-Seq validation on predicted co-regulators

X	*Y* (P/A)	*Z*	CMF motif	Ratio	*P*-value
Klf4	cMyc (A)	Tcf3	TCAGAGa	0.266	7.8e-12
Oct4	cMyc (A)	Sox2	gcaaAACAAAG	0.307	2.4e-20
Sox2	Zfx (P)	nMyc	CACGTCg	5.7	6.0e-45
Stat3	nMyc (A)	Sox2	aaCACAAAGg	0.686	7.3e-3

Note: P/A refers to presence or absence of *Y*. The subsequence of a CMF motif that aligns to the known consensus is given in upper case.

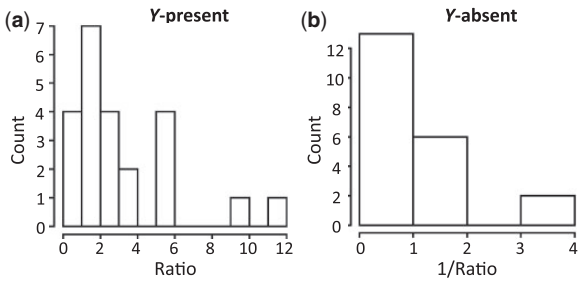

Fig. 3. Validation of predicted CD co-regulators. (**a**) Histogram of the ratio of the occurrence rates of ChIP peaks in S_{XY} over $S_{X\setminus Y}$ for *Y*-present co-regulators. (**b**) Histogram of the inverse ratio for *Y*-absent co-regulators

peak occurrence rate of *Z* in S_{XY} to that in $S_{X\setminus Y}$ is >20 and five triplets with the inverse ratio >5. These are the cases with the strongest evidence from ChIP-Seq data for *Z* being a CD co-bound TF of *X*. The *P*-value of the LA score, $LA(X, Z|Y)$, is <0.01 (our cutoff) for 19 of the 50 triplets. This shows that the LA filtering step is expected to identify ~40% of the CD co-regulators of a TF, which seems satisfactory for a large-scale screening given the high validation rate.

2.4 CD co-regulation network of core TFs

The 27 validated CD co-regulatory interactions among the 14 core TFs are particularly interesting and reliable as supported by ChIP-Seq data, gene expression data and *de novo* motif finding. These predictions may be incorporated into a network that represents CD collaborations among the TFs. In this network, a directed edge from *X* to *Z* means that *Z* is a co-regulator of *X*. An edge from *Y* to the edge between *X* and *Z* indicates that the co-regulatory interaction between *X* and *Z* depends on *Y*. See the legend of Figure 4.

There are eight validated *Y*-absent co-regulators, and for most of them, the binding context is defined by the absence of cMyc and/or nMyc, two Myc family members with functional redundancy and overlapping expression during early development (Malynn *et al.*, 2000). Presented in Figure 4a is the subnetwork of Myc-absent co-regulation between the core factors. Most of the core TFs in this Myc-absent subnetwork are from the Oct4-group TFs defined by Chen *et al.* (2008), including Oct4, Sox2, Nanog, Stat3 and Smad1. They also defined another distinct group of TFs, the cMyc group, which tend to form clusters of

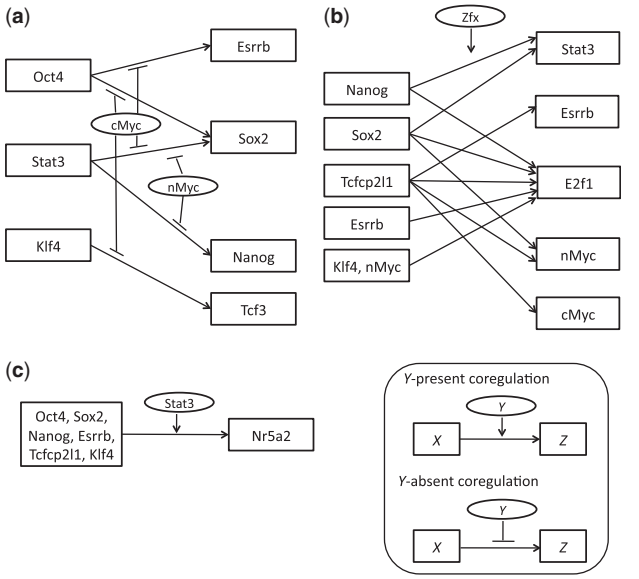

Fig. 4. CD co-regulation networks among the core TFs. (**a**) The Myc-absent co-regulation network. (**b**) The Zfx-present co-regulation network. (**c**) Subnetwork of the co-regulator Nr5a2. See legend for the meaning of different types of edges

binding sites. The cMyc group includes cMyc, nMyc, Zfx and E2f1. Furthermore, Klf4, Esrrb and Tcfcp2l1 showed more frequent co-occupancy with the Oct4 group than with the cMyc group [Fig. 4A in Chen *et al.* (2008)], and therefore, we also classify them into the Oct4 group. The subnetwork in Figure 4a shows two general binding patterns of the Oct4 group regulators. They may either co-occupy regulatory elements with other Oct4 group members or with Myc TFs. For example, Sox2 is a cMyc-absent co-regulator of Oct4, which means that Oct4 binds some regions with Sox2 and other regions with cMyc. See right panel of Figure 1a for illustration. Previous studies have demonstrated the distinct regulatory roles of the two combinatorial binding patterns. In both ESCs and induced pluripotent stem cells (iPSCs), genes associated with cMyc binding, by itself or in combination with the other three reprogramming factors, are significantly enriched for regulators of metabolic processes, while genes co-bound by Oct4, Sox2 and Klf4 in absence of cMyc are mainly implicated in regulation of development (Sridharan *et al.*, 2009). Genes upregulated in ESCs are enriched in the set of genes extensively bound by the Oct4 group factors, while downregulated genes are more likely to be co-bound by both groups (Chen *et al.*, 2008). Simultaneous knockout of cMyc and nMyc in ESCs led to a significant increase in the expression of primitive endoderm markers Gata6 and FoxA2 (Neri *et al.*, 2012; Smith *et al.*, 2010), showing that Myc maintains the undifferentiated state of ESCs by repressing genes involved in early differentiation. Combined with these published data, our result of the Myc-absent co-regulation network suggests that Myc TFs may act like a regulatory switch: When they occupy the binding neighborhood of Oct4 group factors, the overall regulatory role of the binding sites become repressive. Moreover, our result implies that clustering among the Oct4 group binding sites is less tight when cMyc/nMyc binding is present. For example, Sox2 binding

sites are depleted by 70% in the neighborhood of Oct4 sites when cMyc co-occupies the region (second row in Table 2). Consistently, we observed that clustering between the binding sites of Oct4 and Sox2 is significantly stronger ($P = 3.8 \times 10^{-3}$) in the upstream regions of ESC upregulated genes than those of ESC downregulated genes (Cha and Zhou, submitted for publication).

The *Y*-present co-regulation network among the core TFs is more complex with 19 ChIP-Seq validated edges in total. We noticed two significant patterns among these links. First, for 12 of the edges the context is defined by Zfx binding (Fig. 4b). Galan-Caridad *et al.* (2007) reported a critical role of Zfx in promoting ESC self-renewal, although deletion of Zfx only minimally affected the expression of Nanog, Oct4 and Sox2. The exact regulatory function of Zfx binding in ESCs is still unclear. Our analysis suggests a new clue to this question: Zfx binding appears to promote the co-occupancy between and within both groups of core TFs. As such, Zfx may be important for a robust regulation of essential target genes for ESC self-renewal and maintenance by multiple and perhaps redundant TFs. Therefore, deletion of Zfx may weaken the collaboration among these core TFs, and thus impair the overall expression of self-renewal genes but without much effect on the expression of the core TFs such as Oct4, Sox2 and Nanog. As a cMyc group TF, the role of Zfx in co-regulation between Oct4 group factors, Nanog, Sox2 and Stat3 (Fig. 4b), is particularly interesting. We extracted target genes of the co-bound regions of Nanog, Zfx and Stat3 and those of Sox2, Zfx and Stat3. Hereafter, the target gene is defined as the nearest gene to a binding region if the region is within −10 to 2 kb relative to the transcription start site of the gene. The top three significant gene ontology (GO) terms (Gene Ontology Consortium, 2000) enriched in this set of target genes, compared against all mouse genes, are methylation-dependent chromatin silencing ($P = 4.8 \times 10^{-5}$), single-organism developmental process ($P = 3.2 \times 10^{-4}$) and covalent chromatin modification ($P = 3.8 \times 10^{-4}$). The significance level of a GO term was calculated by the online GO tool at go.princeton.edu. As a comparison, the target genes co-bound by Nanog and Stat3 or co-bound by Sox2 and Stat3, but lacking Zfx binding in the nearby 5 kb regions, are mostly enriched for genes relevant to metabolic processes. This suggests that clusters of Oct4 group binding sites with Zfx binding may regulate genes responsible for chromatin modifications, another possible means by which Zfx may promote self-renewal in ESCs.

Second, Nr5a2, a nuclear receptor critical in regulating the expression of Oct4 (Gu *et al.*, 2005), is a CD co-regulator of six core factors given Stat3 binding is present (Fig. 4c). It has been reported recently that Nr5a2 can replace Oct4 in the derivation of iPSCs from mouse somatic cells, and genome-wide binding data showed that Nr5a2 colocalizes with the Oct4 group TFs, such as Oct4, Sox2, Nanog and Esrrb (Heng *et al.*, 2010). Our analysis further revealed that when Stat3 sites are present, the colocalization of Nr5a2 and the aforementioned Oct4 group TFs becomes stronger. As Stat3 is downstream of the LIF pathway, this result highlights the role of the interplay between the LIF pathway and the core regulatory circuit. For every pair of TFs, *X* and *Y* (Stat3), in Figure 4c, we extracted their co-bound regions that contain Nr5a2 peaks and those that lack Nr5a2 peaks in the 5 kb neighborhood. Interestingly, the co-bound regions with Nr5a2 peaks have a much higher percentage of target genes relevant to developmental processes (36%) than those co-bound regions lacking Nr5a2 binding (27%). Because the only difference between the two sets of regions is the presence or absence of Nr5a2 sites, this comparison provides another piece of evidence for the collaborative role of Nr5a2 in regulating developmental genes.

2.5 Predicted CD co-regulators

Now we turn to the top predicted CD co-regulators ranked by the permutation *P*-values of their LA scores. Reported in Table 3 are the ones with *P*-value $\leq 10^{-5}$, together with the core TFs *X* they work with and the TFs *Y* that define the context. Many of the top CD co-regulators are themselves core regulators, such as Oct4, Sox2, Nanog, Esrrb, E2f1 or their close family members or protein-interaction partners, including Sox4, Stat1, Tcfcp2, Zfp238 and Tfdp1 (which forms a dimer with E2f1). This confirms the reliability of our prediction given the high validation rate among the core TFs. At least three stem cell–related GO terms are enriched among these top co-regulators: stem cell maintenance ($P = 5.9 \times 10^{-5}$), stem cell development ($P = 5.9 \times 10^{-5}$) and stem cell differentiation ($P = 2.9 \times 10^{-4}$). The enrichment was calculated against the set of all mouse TFs that have a PWM in TRANSFAC, as our co-regulators were predicted from this background set. The enrichment level of these three GO terms is at least 2- to 3-folds higher than the background set. Together with the many core TFs predicted in the top list, this validates the functional relevance of our predicted co-regulators. Besides those associated with the three mentioned GO terms, some predicted co-regulators have been reported recently in the literature on their regulatory roles in ESCs or early differentiation. We selectively discuss a few of them, which are most relevant to this study.

Grskovic *et al.* (2007) established that the main role of the TF Nfy is to maintain the high proliferative capacity of ESCs. This factor is a trimer composed of three subunits, Nfya, Nfyb and Nfyc, two of which are top co-regulators in our prediction. Using a competition assay (Ivanova *et al.*, 2006), Grskovic *et al.* observed that ESCs infected with Nfya or Nfyb short hairpin RNAs showed a lower growth rate and were selectively outcompeted by wild-type cells. More recently, Dolfini *et al.* (2012) demonstrated the activation role of Nfy on key ESC regulatory genes, including Sox2, Oct4, Nanog, Klf4, Sall4 and Jarid2, by using dominant negative mutant and transduced overexpression of Nfya, which encodes the DNA-binding subunit of the TF. They also confirmed that Nfy binds the promoters of these genes. By comparing ChIP-Seq data of Nfya (Tiwari *et al.*, 2011) and those of the core regulators in ESCs (Chen *et al.*, 2008), significant co-occupancy of promoters was observed between this TF and Oct4, Stat3 and Nanog with *P*-values on the order of 10^{-12}, 10^{-7} and 10^{-50}, respectively [Fig. 5A in Dolfini *et al.* (2012)]. Our method further predicted a significant collaboration between Nfy and a few key regulators, Oct4, Stat3 (Table 3) and Nanog (LA $P = 2 \times 10^{-5}$, Supplemental Table S1), when cMyc (or Oct4) binding is absent. In light of the switcher role of cMyc discussed above (Fig. 4a), our prediction suggests that Nfy

Table 3. Top predicted CD co-regulators and enriched GO terms

X	*Y* (P/A)	*Z* (GO terms)
Oct4	cMyc (A)	Esrrb (1,2,3), Sox2 (1,2,3), Nfya, Foxh1, Irf9
Oct4	Stat3 (A)	Ets1, Foxa3, Hsf1, Sfpi1 (1,2,3)
cMyc	Zfx (P)	Srebf1
Esrrb	Zfx (P)	Trp53
Nanog	Oct4 (A)	E2f1, Tfdp1
nMyc	Tcfcp2l1 (A)	E2f1
Smad1	Nr5a2 (A)	Elf1
Smad1	Zfx (P)	Nr2f2
Sox2	Zfx (P)	Tfdp1
Stat3	Oct4 (A)	Ar, Bach1, Ets1, Ets2, Foxf1a, Gabpb1, Gli1, Irf5, Nfkb2, Nfyb, Pax8 (1,2,3), Pitx2, Rfxank, Sox4 (1,2,3), Srebf1, Stat1, Tcfcp2, Tfdp1, Usf1, Usf2
Stat3	cMyc (A)	Rxra, Sox18 (3), Sox2 (1,2,3)
Stat3	nMyc (A)	Cebpa, Hmga2 (1,2,3), Irf9, Nanog (1,2,3), Rbpj (1,2,3), Sox2 (1,2,3)
Stat3	Tcfcp2l1 (A)	Cebpa, Oct4 (1,2,3), Trp53
Stat3	Zfx (P)	Bhlhb2, Srebf1, Tfdp1
Tcf3	Oct4 (A)	Bhlhb2, Zfp238
Tcfcp2l1	Zfx (P)	E2f1, Maz, Nanog (1,2,3), Rela, Rest (3)

Note: Numbers in the parentheses indicate which, if any, of the following three GO terms are associated with the co-regulator: (1) stem cell maintenance, (2) stem cell development and (3) stem cell differentiation.

is a coherent component of the core regulatory circuit that regulates pluripotent genes in ESCs.

We see a clear enrichment of nuclear receptors, Esrrb (Nr3b2), Nr2f2, Ar (Nr3c4) and Rxra (Nr2b1), in this top list of predicted co-regulators. As mentioned above, another nuclear receptor Nr5a2 is a prominent co-regulator identified in the core co-regulatory network (Fig. 4c). Because Dax1 (Nr0b1) forms a complex with Nr5a2 and is recruited to the binding sites of Nr5a2 (Kelly *et al.*, 2010), it is possible that the prominent co-regulatory role of Nr5a2 comes partly from the involvement of Dax1, a protein-interaction partner of Nanog (Wang *et al.*, 2006). Moreover, another predicted co-regulator, Trp53, is upregulated by both Nr5a2 and Dax1 in mouse ESCs (Kelly *et al.*, 2010). This long list of nuclear receptors in our prediction is in line with the many recent studies that establish the regulatory role of Esrrb, Nr5a2 and Dax1 on the expression and activity of key TFs such as Oct4, Nanog or Klf4, and vice versa. See Jeong and Mangelsdorf (2009) for a review. Three nuclear receptors, Nr2f2, Ar and Rxra, have not been implicated in mouse ESCs, although some of them have established roles in the differentiation of adult mesenchymal stem cells and neural stem cells (Jeong and Mangelsdorf, 2009). Interestingly, these three factors are predicted as co-regulators of Smad1 or Stat3, which are respective downstream factors of the BMP and the LIF signaling pathways. This result suggests possible interactions between these nuclear receptors and the two pathways that are central to the maintenance of a pluripotential stem cell phenotype.

We predicted 20 CD co-regulators of Stat3 when Oct4 binding is absent, a much larger number than those of the other pairs in Table 3. All these co-regulator genes have the most significant LA scores, but some of them have no established functions in ESCs. How Stat3 interacts with these potential co-regulators awaits future experimental clarifications.

3 DISCUSSION

We used a computational method to predict CD co-regulators based on ChIP-Seq data of two TFs and gene expression data. This method can be applied to many available ChIP-Seq and gene expression data, and is able to decode complex combinatorial regulation among multiple regulators. Our application to mouse ESC data demonstrated the novel biological insights that can be generated by this method. In particular, we constructed a reliable CD co-regulation network among 14 core TFs in mouse ESCs and identified a list of novel co-regulators for future investigation.

A few lines of future work may further widen the application of our approach. It is tempting to detect co-regulators under more complex binding context, say, defined by the combinatorial binding of three or more TFs. However, an efficient way to identify a small set of most relevant binding contexts is crucial, as a naive comparison among all combinatorial binding regions will be prohibitive computationally. Therefore, a principled statistical method to detect combinatorial binding regions from ChIP-Seq data of multiple (potentially a large collection of) TFs is highly desired. Generalization of CMF to an efficient tool that finds differentially enriched motifs by comparing multiple sets of sequences is another necessary step. It is also interesting to study the relation between epigenetic factors and CD binding of TFs.

4 METHODS

4.1 Preprocessing of ChIP-Seq data

Genomic coordinates of ChIP-Seq peaks of the 12 TFs in Chen *et al.* (2008) were downloaded from their supplemental materials. For Tcf3 and Nr5a2, we used the coordinates provided by the database hmChIP (Chen *et al.*, 2011). For each TF, we expanded 200 bp on both sides of a peak

and recursively merged overlapping expanded peaks. Then, we constructed two sets of sequences for an ordered pair X and Y using two stringent distance cutoffs. Peaks (after the above preprocessing) of X that lack Y peaks within 5 kb were used to construct $S_{X\setminus Y}$. On the other hand, we recursively merged two peaks if their distance is <50 bp and if the merged region contains at least one peak from each of the two TFs. The merged regions are defined as the co-bound sequence set S_{XY}, which is identical to S_{YX}. Then the DNA sequence of a binding region was extracted by the software CisGenome (Ji *et al.*, 2008).

4.2 Matching motifs to known PWMs

We first retrieved the consensus pattern from a PWM. The dominating nucleotide of a position is defined as the consensus if it has at least 50% of the observed counts; otherwise the consensus of this position is defined as an 'N'. We discarded a PWM if its consensus sequence contains more than two 'N's in every subsequence of length seven. The (length-seven) subsequence of which the dominating nucleotides have the highest total percentage of counts is defined as the consensus pattern. Then we calculated the number of matches between the consensus pattern of a PWM in TRANSFAC and every sliding window of length seven along a PF motif. If there are at least five matches (excluding 'N') in a sliding window, then we say that the motif matches the PWM.

4.3 Liquid association

LA is a generalized notion of association for describing certain kind of ternary relationship among variables. Suppose that X, Y and Z are three random variables with mean zero and unit variance. Let $g(Y) = E(XZ|Y)$ be the conditional correlation between X and Z given Y. The LA of X and Z with respect to Y is defined by

$$LA(X, Z|Y) = E[g'(Y)] \tag{2}$$

which measures the average rate of change in the conditional correlation. By Stein's Lemma, if Y is normally distributed,

$$E[g'(Y)] = E[g(Y)Y] = E[E(XZY|Y)] = E(XZY), \tag{3}$$

which leads to the estimation (1) when we have observations $\mathbf{x}, \mathbf{y}, \mathbf{z}$ for the three random variables.

Funding: This work was supported by NSF grant DMS-1055286 to Q.Z.

Conflict of Interest: none declared.

REFERENCES

Bailey,T.L. (2011) DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, **27**, 1653–1659.

Chen,G. and Zhou,Q. (2011) Searching ChIP-Seq genomic islands for combinatorial regulatory codes in mouse embryonic stem cells. *BMC Genomics*, **12**, 515.

Chen,L. *et al.* (2011) hmChIP: a database and web server for exploring publicly available human and mouse ChIP-seq and ChIP-chip data. *Bioinformatics*, **27**, 1447–1448.

Chen,X. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem. *Cell*, **133**, 1106–1117.

Dolfini,D. *et al.* (2012) The short isoform of NF-YA belongs to the embryonic stem cell transcription factor circuitry. *Stem Cells*, **30**, 2450–2459.

Galan-Caridad,J.M. *et al.* (2007) Zfx controls the self-renewal of embryonic and hematopoietic stem cells. *Cell*, **129**, 345–357.

Gene Ontology Consortium. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Grskovic,M. *et al.* (2007) Systematic identification of cis-regulatory sequences active in mouse and human embryonic stem cells. *PLoS Genet.*, **3**, e145.

Gu,P. *et al.* (2005) Orphan nuclear receptor LRH-1 is required to maintain Oct4 expression at the epiblast stage of embryonic development. *Mol. Cell. Biol.*, **25**, 3492–3505.

Heng,J.C. *et al.* (2010) The nuclear receptor Nr5a2 can replace Oct4 in the reprogramming of murine somatic cells to pluripotent cells. *Cell Stem Cell*, **6**, 167–174.

He,X. *et al.* (2009) A biophysical model for analysis of transcription factor interaction and binding site arrangement from genome-wide binding data. *PLoS One*, **4**, e8155.

Ivanova,N. *et al.* (2006) Dissecting self-renewal in stem cells with RNA interference. *Nature*, **442**, 533–538.

Jeong,Y. and Mangelsdorf,D.J. (2009) Nuclear receptor regulation of stemness and stem cell differentiation. *Exp. Mol. Med.*, **41**, 525–537.

Ji,H.K. *et al.* (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.*, **26**, 1293–1300.

Kelly,V.R. *et al.* (2010) Dax1 up-regulates Oct4 expression in mouse embryonic stem cells via LRH-1 and SRA. *Mol. Endocrinol.*, **24**, 2281–2291.

Kim,J. *et al.* (2008) An extended transcriptional network for pluripotency of embryonic stem cells. *Cell*, **132**, 1049–1061.

Li,K. (2002) Genome-wide coexpression dynamics: Theory and application. *PNAS*, **99**, 16875–16880.

Malynn,B.A. *et al.* (2000) N-myc can functionally replace c-myc in murine development, cellular growth, and differentiation. *Genes Dev.*, **14**, 1390–1399.

Marson,A. *et al.* (2008) Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell*, **134**, 521–533.

Mason,M. *et al.* (2010) Identification of context-dependent Motifs by contrasting ChIP binding data. *Bioinformatics*, **26**, 2826–2832.

Matys,V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.

Neri,F. *et al.* (2012) Myc regulates the transcription of the PRC2 gene to control the expression of developmental genes in embryonic stem cells. *Mol. Cell. Biol.*, **32**, 840–851.

Niwa,H. *et al.* (1998) Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3. *Gene Dev.*, **12**, 2048–2060.

Smith,K.N. *et al.* (2010) Myc represses primitive endoderm differentiation in pluripotent stem cells. *Cell Stem Cell*, **7**, 343–354.

Sridharan,R. *et al.* (2009) Role of the murine reprogramming factors in the induction of pluripotency. *Cell*, **136**, 364–377.

Takahashi,K. and Yamanaka,S. (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factor. *Cell*, **126**, 663–676.

Tiwari,V.K. *et al.* (2011) A chromatin-modifying function of JNK during stem cell differentiation. *Nat. Genet.*, **44**, 94–100.

Thomas-Chollier,M. *et al.* (2012) RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res.*, **40**, e31.

Wang,J. *et al.* (2006) A protein interaction network for pluripotency of embryonic stem cells. *Nature*, **444**, 364–368.

Zhou,Q. *et al.* (2007) A gene regulatory network in mouse embryonic stem cells. *PNAS*, **104**, 16438–16443.