Interactions between pluripotency factors specify *cis*-regulation in embryonic stem cells

Chris Fiore and Barak A. Cohen

Center for Genome Sciences and Systems Biology, Department of Genetics, Washington University School of Medicine, St. Louis, Missouri 63110, USA

We investigated how interactions between pluripotency transcription factors (TFs) affect *cis*-regulation. We created hundreds of synthetic *cis*-regulatory elements (CREs) comprised of combinations of binding sites for pluripotency TFs and measured their expression in mouse embryonic stem (ES) cells. A thermodynamic model that incorporates interactions between TFs explains a large portion (72%) of the variance in expression of these CREs. These interactions include three favorable heterotypic interactions between TFs. The model also predicts an unfavorable homotypic interaction between TFs, helping to explain the observation that homotypic chains of binding sites express at low levels. We further investigated the expression driven by CREs comprised of homotypic chains of KLF4 binding sites. Our results suggest that KLF homologs make unique contributions to regulation by these CREs. We conclude that a specific set of interactions between pluripotency TFs plays a large role in setting the levels of expression driven by CREs in ES cells.

[Supplemental material is available for this article.]

Sequence-specific transcription factors (TFs) direct the *cis*-regulatory programs that govern mammalian development. Two competing models describe the combined action of TFs at *cis*-regulatory elements (CREs), the "enhancesome model" and the "billboard model." In the enhancesome model, TFs interact with each other in ways that are exquisitely sensitive to the spatial arrangement of other bound TFs (Thanos and Maniatis 1995). In contrast, the billboard model states that TFs act independently and are relatively insensitive to their surrounding DNA context (Kulkarni and Arnosti 2003; Arnosti and Kulkarni 2005). Support for the enhancesome model comes from several studies that demonstrate orientation-dependent TF–TF interactions (Thanos and Maniatis 1995; Kim and Maniatis 1997; Senger et al. 2004; Panne et al. 2007; Gertz et al. 2009; Goldwater et al. 2010; Sharon et al. 2012; Yáñez-Cuna et al. 2012; Smith et al. 2013; Erceg et al. 2014). However, counter examples in which TFs regulate transcription independent of the specific arrangement of binding sites support a more billboard-like model (Arnosti et al. 1996; Kulkarni and Arnosti 2003; Liu and Posakony 2012). Thus, it remains a challenge to determine whether specific TFs interact with each other, and if so, to understand whether these interactions constrain the arrangements of their cognate binding sites in regulatory DNA.

The regulation of pluripotency in embryonic stem (ES) cells is an important model system for studying TF interactions. Pluripotency is maintained by a core set of TFs, which include POU5F1 (also known as OCT4), SOX2, KLF2, KLF4, KLF5, MYC, NANOG, and ESRRB (Boyer et al. 2005; Ivanova et al. 2006; Loh et al. 2006; Chen et al. 2008b). Evidence suggests that these pluripotency TFs act cooperatively in ES cells. POU5F1 and SOX2 physically interact to regulate several target genes (Chew et al. 2005; Kuroda et al. 2005; Rodda et al. 2005). Chromatin immunoprecipitation (ChIP) studies also show that groups of pluripotency TFs often bind in clusters at common genomic loci (Chen et al. 2008b; Kim et al. 2008). Thus, it is likely that interactions between pluripotency TFs play a role in the pluripotency network, but the "rules," if any, by which these TFs interact remain obscure.

Synthetic CREs are powerful tools for uncovering interactions between TFs (Gertz et al. 2009; Sharon et al. 2012; Mogno et al. 2013; Smith et al. 2013; Erceg et al. 2014; Levo et al. 2015). When combined with massively parallel reporter gene assays (Kinney et al. 2010; Kwasnieski et al. 2012; Melnikov et al. 2012; Patwardhan et al. 2012; Sharon et al. 2012; Arnold et al. 2013), libraries of rationally designed synthetic regulatory elements provide the statistical power to uncover *cis*-regulatory interactions. Here, we used libraries of synthetic CREs to study the interactions between TFs that maintain pluripotency in ES cells. Our results help explain how interactions between pluripotency TFs affect transcriptional regulation in ES cells.

Results

Expression of synthetic CRE library in ES cells

We designed a synthetic CRE library to investigate combinatorial *cis*-regulation by pluripotency TFs. Our goal was to test two possible *cis*-regulatory scenarios—one in which TFs act independently and one in which interactions between TFs guide regulation. To distinguish between these scenarios, we designed synthetic CREs with varying combinations of binding sites for pluripotency TFs and then measured their expression in ES cells. Each reporter construct consisted of a synthetic CRE with between one and four binding sites for pluripotency TFs upstream of the *Pou5f1* basal promoter fused to the dsRed reporter gene and a unique sequence barcode in the 3′ UTR of the dsRed gene (Supplemental Fig. S1). The synthetic CRE was placed immediately upstream of the basal promoter, as done previously (White et al. 2013; Kwasnieski et al. 2014), to reduce the chance that spurious binding sites in

Corresponding author: cohen@genetics.wustl.edu
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.200733.115.

intervening spacer sequences influence the expression of our constructs (Gertz et al. 2009). This experimental design maximizes the chances of capturing the effects of intrinsic interactions between TFs on expression. The binding sites in the synthetic CREs are high-affinity binding sites for four of the TFs central to pluripotency: POU5F1, SOX2, KLF4, and ESRRB. The lack of a well-defined binding site for NANOG precluded us from including a NANOG binding site in our synthetic CREs. Each binding site resides in a 20-bp sequence where it is surrounded by at least 8 bp of constant sequence (Supplemental Fig. S1). This spacing (20 bp) places neighboring binding sites on the same side of the DNA helix. The unique sequence barcodes in the 3′ UTRs of the constructs allowed us to use a massively parallel reporter gene assay to measure the activity of hundreds of constructs in parallel (Kwasnieski et al. 2012).

We built a library (OSKE library) of 599 synthetic CREs composed of different numbers and combinations of the binding sites for POU5F1, SOX2, KLF4, and ESRRB. To provide redundancy in our measurements, each synthetic CRE is in 10 different constructs, each of which has a unique sequence barcode (BC). We transfected the OSKE library into mouse ES cells and measured the expression of the CREs 26 h later by sequencing the barcodes in RNA extracted from transfected cells. After filtering for quality control (Methods), expression data were obtained for 3567 BCs corresponding to 415 unique synthetic regulatory elements. The expression measurements of these CREs were highly reproducible across three biological replicates (R^2 ranged from 0.88 to 0.91 between replicates) (Supplemental Fig. S1C). CREs with more binding sites did tend to express higher, but this trend explains only a small portion of the variation in expression ($R^2 = 0.14$). Furthermore, expression among the subset of CREs with exactly four TF binding sites (TFBS) varied over a 13-fold range, suggesting that differences between the different TFBS, as well as interactions between TFs, account for much of the difference in activity among CREs.

Thermodynamic model of OSKE expression

To investigate whether interactions between TFs might explain the trends in our library, we analyzed expression using a thermodynamic model (Shea and Ackers 1985; Buchler et al. 2003; Bintu et al. 2005; Segal et al. 2008; Gertz et al. 2009; He et al. 2010; Kinney et al. 2010; Sherman and Cohen 2012; Brewster et al. 2014; Zeigler and Cohen 2014). The model provides a formal biophysical framework to represent TF–TF interactions and their effects on gene expression. In the model, a regulatory element is described as a collection of thermodynamic states in which each individual site is either bound or unbound by its cognate TF. Bound TFs either promote or inhibit the recruitment of RNA polymerase, and the probability that RNA polymerase is bound at a CRE is taken as being proportional to the output expression of that CRE. To compute this probability, the model uses parameters that describe the free energies of TF–TF and TF–RNAP interactions on CREs (Sherman and Cohen 2012). We fit the model by finding a set of parameters for these interactions that yields the best match to the measured expression levels across all library members. Our goal was to determine which parameters (TF interactions) were necessary to accurately describe the differences in expression we observed in the OSKE library.

We first fit a thermodynamic model including only four TF–RNAP interaction parameters, one for each TF, but allowing no TF–TF interaction parameters (Fig. 1A). This is analogous to fitting a linear regression with only main effects for each TF. The advantage of the thermodynamic model over linear regression is that the thermodynamic model naturally captures biophysical nonlinearities, such as binding site saturation, without extra parameters. With linear regression, modeling binding site saturation requires higher order "interaction" terms, even when all sites act independently. We found that the thermodynamic model, in which each TF contributes independently, explained 50% of the total variance in expression in the library (Fig. 1A). This result suggests that a large fraction of the differences between different library members comes from the independent action of the four TFs.

We next asked whether interactions between TFs might also contribute to differences in the expression levels of different library members. To do this, we compared the thermodynamic model above (with TFs acting independently) to a model in which we included parameters for TF–TF interactions (akin to linear regression with higher order interactions). In this analysis, we allowed two types of rules for TF–TF interactions: the "neighboring" interactions rule, in which TFs interact with each other only

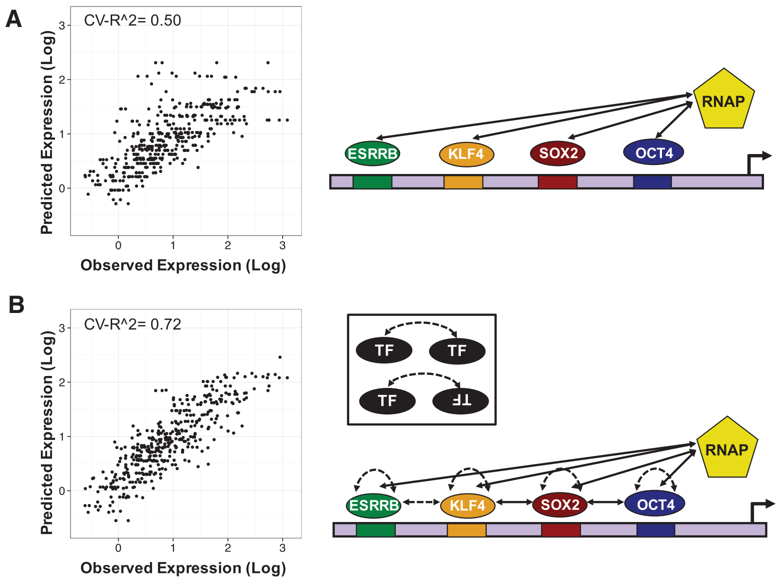

Figure 1. Thermodynamic model of OSKE library. In scatter plots, observed expression of each CRE from massively parallel reporter assays (Supplemental Fig. S1; Supplemental Data S1) is plotted on the *x*-axis, and the predicted expression of each CRE by the model is on the *y*-axis. In depictions of models, solid lines represent interactions following the "neighboring" rule, and dashed lines represent interactions following the "across" rule (Supplemental Fig. S2). See Table 1 for parameter values. (*A*) Model with only four TF–RNAP interaction parameters predicts expression with cross-validated R^2 of 0.50. (*B*) Full model with five TF–TF interaction parameters in addition to four TF–RNAP interaction parameters predicts expression with cross-validated R^2 of 0.72. Interactions were validated using genomic data (Supplemental Fig. S3).

when bound at adjacent sites, and the "across" rule, in which TFs interact with each other independent of their spacing or the presence of other bound proteins in between the sites (Methods; Supplemental Fig. S2). By comparing model variants, we examined the explanatory power of possible TF–TF interactions.

The best model with TF–TF interactions ("Full Model") includes nine total parameters and explains expression in the library with an R^2 of 0.72 (Fig. 1B). To guard against overfitting, we monitored the Akaike Information Criterion (AIC) (Akaike 1974) and sensitivity of the parameter values and used fivefold cross validation. The Full Model includes the four independent TF–RNAP parameters, two homotypic TF–TF interaction parameters, as well as three of the six possible heterotypic TF–TF interactions: SOX2-POU5F1, KLF4-SOX2, KLF4-ESRRB (Table 1). The interaction between SOX2 and POU5F1 is well documented and validates the ability of our model to detect known interactions (Ambrosetti et al. 2000; Chew et al. 2005; Kuroda et al. 2005; Rodda et al. 2005). Analysis of genomic ChIP-seq binding data (Chen et al. 2008b) for the KLF4-ESRRB pair and the KLF4-SOX2 pair provides support that these TFs also bind cooperatively in the genome (Supplemental Fig. S3). Taken together, these data suggest that some TFs contribute independently to expression while certain pairs of TFs, including KLF4-ESRRB, KLF4-SOX2, and POU5F1-SOX2, interact cooperatively to set the activity of CREs. The thermodynamic framework quantitatively models these effects and supports the conclusion that cooperativity between different TFs helps determine the expression driven by CREs in the pluripotency network.

In the pluripotency system, the nature of the heterotypic interactions between TFs excludes a strict version of the "billboard" model in which the arrangement of binding sites does not matter. The KLF4-SOX2 and SOX2-POU5F1 interactions both use the neighboring interactions rule, constraining the highest expressing CREs to configurations in which these sites are directly adjacent to one another (see Methods). Furthermore, the SOX2-POU5F1 interaction has an orientation dependence that requires that POU5F1 be closer to the transcription start site. In contrast, the KLF4-ESRRB interaction is order and orientation independent. Each of these constraints is supported by a comparison between models with and without the constraints (see Methods). The results demonstrate that for a CRE with binding sites for multiple TFs, the order in which the binding sites are arranged can have an effect on the strength of TF interactions, and thus on the output expression.

Table 1. Fit parameter values from thermodynamic models

Parameter	Value	95% C.I.
OSKE library		
ESRRB-RNAP	0.8915	0.7788, 1.004
KLF4-RNAP	1.06	0.9294, 1.19
POU5F1-RNAP	0.5807	0.4729, 0.6884
SOX2-RNAP	0.366	0.2515, 0.4805
Homotypic same orientation (A)	−1.132	−1.476, −0.7874
Homotypic opposite orientation (A)	−2.468	−3.29, −1.65
KLF4-ESRRB (A)	1.119	0.739, 1.499
POU5F1-SOX2, only with POU5F1 closer to TSS (N)	0.981	0.1595, 1.802
KLF4-SOX2 (N)	1.336	0.8047, 1.867
KBS library		
KLF2-RNAP	0.1593	−0.2326, 0.5511
KLF4-RNAP	0.2306	−0.05295, 0.5142
KLF5-RNAP	2.169	1.063, 3.274
KLF4-KLF4 (N)	2.369	1.344, 3.393

(N) indicates an interaction with the neighboring interactions rule; (A) indicates an interaction with an all across interaction rule. Positive values are favorable, and negative values are unfavorable.

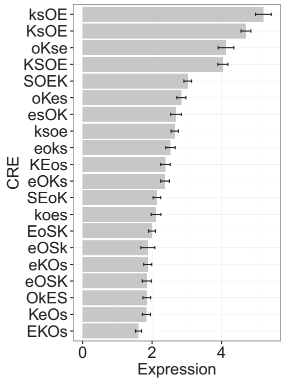

Figure 2. Expression of CREs with four unique binding sites. Expression of the 20 CREs with one binding site for each of the four TFs in the OSKE library. In CRE names: (O) POU5F1 (OCT4) binding sites; (S) SOX2 binding sites; (K) KLF4 binding sites; and (E) ESRRB binding sites. Lowercase letters indicate the reverse-orientation. Expression represented as mean ± SEM.

Comparisons between CREs with different arrangements of the same binding sites highlight the importance of TF interactions. There are 20 CREs in the library with exactly one binding site for each of the four TFs (Fig. 2). These CREs vary only in the relative arrangement of the binding sites. Under a strict Billboard model, these CREs should all drive the same level of expression. In contrast to this prediction, we observe that expression from these CREs varies over a threefold range. Application of the rules we deduced from the entire library helps explain the differences between these 20 CREs. The CREs with the four highest expression values all contain adjacent SOX2 and KLF4 sites, and all but one have adjacent SOX2 and POU5F1 sites with POU5F1 closer to the basal promoter. The thermodynamic model explains 36% of the variation in expression among these 20 CREs, relying only on those two interactions. These results show that interactions following the neighboring interactions rule explain a significant fraction of the expression differences between CREs varying only in the arrangement of their binding sites.

Repressive homotypic interactions

Two homotypic TF interactions, defined as interactions between any TF and another copy of itself, also contribute to expression. Two types of homotypic interactions were statistically supported in the Full Model, in which the interacting binding sites were on the same strand or on the opposite strand (Fig. 1B; Table 1). Notably, both homotypic interaction parameters are strongly unfavorable, with some of the largest parameter values in the model (Table 1). The unfavorable homotypic interaction parameters suggest that heterotypic CREs, those CREs composed of binding sites for different TFs, will drive higher expression than homotypic CREs, those CREs comprised of multiple copies of the same binding site. Indeed, among synthetic CREs with exactly four total TFBS, higher expression is strongly associated with the number of unique types of binding sites in the CRE (Fig. 3). In other words, CREs with one binding site for each of the four TFs drive much higher expression than CREs with four copies of a single site. In addition, of those CREs with binding sites for two TFs and four

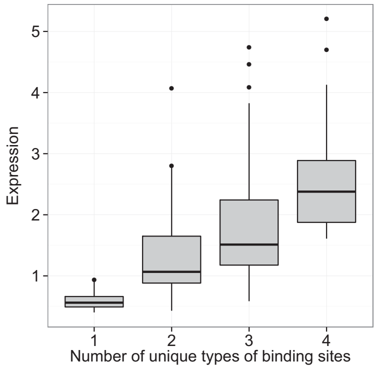

Figure 3. Expression by unique types of binding sites. Expression of CREs in the OSKE library with four total TFBS by the number of unique types of binding sites in the CRE (number of TFs for which binding sites are represented). Data are represented as box plots. Within the group with two unique types of binding sites, CREs with two sites of each type have higher expression than CREs with three sites of one type and one site of the other type (Supplemental Fig. S4).

total binding sites, the expression is higher when there are two binding sites for each TF than when there are three sites for one TF and one site for the other TF (Supplemental Fig. S4). Although these and other results (Smith et al. 2013) show that reduced activity of homotypic chains relative to heterotypic chains plays a role in *cis*-regulation, the mechanism by which this occurs is unclear.

Expression of homotypic chains of KLF4 binding sites

We hypothesized that the expression driven by homotypic chains of TFBS is influenced by competition and interactions between homologs that bind the same binding site. To test this hypothesis, we assayed the expression of CREs comprised of homotypic chains of binding sites for KLF4, the site most associated with activation in the original CRE library. KLF2, KLF4, and KLF5 all regulate pluripotency in ES cells and are known to bind to the KLF4 binding site (Jiang et al. 2008). We created a small CRE library (KBS library) with seven synthetic CREs, comprised of between zero and six KLF4 sites. We then measured the expression of the KBS library in ES cells in which we overexpressed either one of the three *Klf* paralogs or a GFP control.

The expression profile of the KBS library was different in each TF overexpression condition compared to the control, showing that the overexpression of different TFs has different effects on homotypic chains of KLF4 sites (Fig. 4). CREs with three or more KLF binding sites drive higher expression when *Klf2* is overexpressed relative to the control condition ($P < 10^{-4}$, Student's *t*-test). In contrast, *Klf4* or *Klf5* overexpression leads to lower expression from CREs with five or six binding sites compared to the control ($P < 10^{-5}$, Student's *t*-test). Furthermore, when *Klf5* is overexpressed, a CRE with six binding sites drives lower expression than the five-binding-site CRE, although the magnitude of this effect is small ($P = 0.012$, Student's *t*-test). The results show that each KLF homolog has a unique effect on the expression of these CREs. Because each of these TFs is active in ES cells, these results suggest that competition between homologs is an important factor in regulation by these binding sites.

We then used the thermodynamic framework to investigate whether competition and interactions between KLF homologs might explain the observed patterns of expression from homotypic chains of KLF4 binding sites. To parameterize the model, we measured the mRNA expression levels of each *Klf* gene in each overexpression condition by qPCR. Each TF had higher expression in its overexpression condition (Supplemental Fig. S5); however, we also observed cross-regulation between *Klf4* and *Klf5* homologs: *Klf4* was up-regulated in the *Klf5* overexpression condition. We used the measured mRNA levels of the *Klf* homologs to constrain the fits of the model to the overexpression data. We attempted to fit a model using both homotypic and heterotypic interactions between KLF homologs. We again monitored the AIC and parameter sensitivity to guard against overfitting. The model that best explained the data used four parameters: one for each TF–RNAP interaction and a KLF4 homotypic interaction parameter (Table 1). This model captures most of the overall variation in expression ($R^2 = 0.93$) between CREs with different numbers of KLF4 sites and explains the data significantly better than a model with no TF–TF interactions (by AIC) (Supplemental Fig. S6). Although all interactions were favorable, the Klf2-RNAP and KLF4-RNAP interactions were very weak (Table 1). This model suggests that KLF4 is a weak activator with strong self-cooperativity that prevents KLF2 and KLF5 from binding, especially on longer chains of binding sites, and when it is highly expressed. In this model, *Klf4* overexpression leads to lower expression because cooperativity between KLF4 outcompetes the stronger activators for binding. Since the *Klf2* over-expression condition is the only overexpression condition in which *Klf4* is not overexpressed, the expression driven by the CREs with multiple binding sites is highest in this condition because KLF4 cannot dominate the binding.

Finally, we investigated whether clusters of KLF4 binding sites in the genome are associated with activity and might be regulated by a similar mechanism as our synthetic CREs. We used a sliding window approach to investigate the relationship between the number of KLF4 binding motifs in a window and biochemical activity. We found a positive trend between the number of KLF4

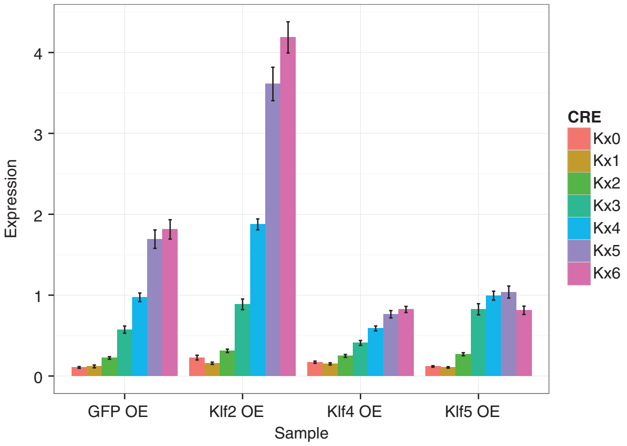

Figure 4. Expression of CREs with only KLF4 binding sites in overexpression backgrounds. Expression of the basal promoter and CREs with one to six KLF4 binding sites in each of the four overexpression (OE) conditions. Expression represented as mean ± SEM (Supplemental Data S1). See text for relevant statistical comparisons. Expression of the *Klf* genes in the OE conditions can be found in Supplemental Figure S5. Notably, *Klf4* and *Klf5* are both overexpressed in the *Klf5* overexpression condition. The expression predictions of the thermodynamic model can be found in Supplemental Figure S6. Biochemical signal of clusters of KLF4 sites in the genome can be found in Supplemental Figure S7.

binding sites and both DNase hypersensitivity signal and RNA Polymerase II binding signal, up to six binding sites, especially within 10 kb of a transcription start site (Supplemental Fig. S7). This effect is greater than would be expected from a set of generic homotypic sequence motifs because repeating the same analysis with permuted KLF4 binding matrices failed to show the same correlation. This also shows that this effect is not due to the high GC-content of the KLF4 binding site because permuted KLF4 binding matrices have the same GC-content. This result demonstrates that genomic clusters of KLF4 binding sites have biochemical activity that is associated with the number of binding sites. Thus, these clusters may regulate transcription in a similar manner as the synthetic CREs, namely through competition and interactions among the KLF factors.

Discussion

Here, we describe an investigation into the *cis*-regulatory activity of TFs in the pluripotency network. A thermodynamic treatment of our data suggests that interactions between TFs play a large role in explaining the expression driven by synthetic CREs, as a model including interactions explains 22% more of the variation in expression than a model without interactions. We further characterized regulation by homotypic chains of one particular binding site, KLF4, and found that KLF2, KLF4, and KLF5 each have unique effects on the activity of these sites.

Our results show that there is *cis*-regulatory logic in the pluripotency network. The expression driven by synthetic CREs with exactly one binding site for each of the four TFs varies over a threefold range. Because all of these CREs have the exact same composition of binding sites, expression differences between these CREs must be due to differences in the order and orientation of the sites relative to each other and to the transcription start site. Consistent with this hypothesis, we found that two of the TF–TF interaction rules in our model, KLF4-SOX2 and SOX2-POU5F1, have a constraint based on the arrangement of the sites, and that the KLF4-SOX2 interaction has an additional order constraint relative to the TSS. These constraints allow the model to capture 36% of the variation in expression between CREs with different arrangements of the same binding sites. The fact that the model cannot capture all of the differences between these CREs suggests that there are additional constraints in this system related to the orientation and arrangement of binding that have yet to be discovered. Libraries with much larger numbers of heterotypic CREs will help uncover these constraints. Further, the synthetic CREs in this system used a specific spacing (20 bp between the beginning of the binding sites) to allow for both binding sites to be on the same side of the DNA helix. However, we recognize that other spacings could result in different TF interactions, such as the SOX2-ESRRB interaction found with a shorter spacing (Hutchins et al. 2013).

TF–TF interaction parameters make important contributions to the explanatory power of the thermodynamic model. These interaction parameters include three favorable heterotypic interactions and two strong unfavorable homotypic interactions. The model found a well-known and characterized interaction between POU5F1 and SOX2 (Chew et al. 2005; Kuroda et al. 2005; Rodda et al. 2005). There is some previous evidence for the KLF4-SOX2 interaction (Nakatake et al. 2006; Wei et al. 2009), and although the KLF4-ESRRB interaction has not been extensively described, Hutchins et al. (2013) did see an enrichment of these motifs co-occurring under ChIP-seq peaks (Hutchins et al. 2013). Furthermore, analysis of genomic binding data supports the KLF4-SOX2 and KLF4-ESRRB interactions (Supplemental Fig. S3). The activities of the synthetic CREs show that these interactions are likely to be important in determining the expression driven by pluripotency TFs. Our work provides a framework demonstrating the quantitative contribution of these interactions to expression.

Our work builds on previous studies showing that homotypic clusters of TF binding sites have unique *cis*-regulatory properties. Generally, homotypic clusters are evolutionarily conserved and located in predicted regulatory regions (Gotea et al. 2010). In one study (Sharon et al. 2012), several homotypic clusters of TF binding sites saturated expression at high levels, but for other sites, homotypic clusters showed reduced expression relative to constructs with fewer sites. Another study also showed that homotypic clusters are associated with a reduced ability to drive expression compared to heterotypic clusters (Smith et al. 2013). Our results also show that homotypic chains of binding sites drive lower expression than heterotypic chains and build on these results with a mechanistic model that captures this quantitative relationship (Fig. 3). Our results support the finding that ES cell promoters that are bound by only one pluripotency TF tend to be off, and those that are bound by multiple TFs tend to be on (Kim et al. 2008). However, homotypic chains of KLF4 binding sites from the KBS library drove higher expression than homotypic chains of other binding sites from the OSKE library. This difference might be explained by the fact that these CREs are regulated by multiple KLF factors (KLF2, KLF4, and KLF5), each of which has a unique effect on expression. Although our thermodynamic modeling can explain most of the variation in the expression driven by homotypic chains of KLF4 binding sites, it cannot capture all of the trends in the expression (Supplemental Fig. S6). For instance, it was unable to capture the saturation in expression between five and six binding sites. These results combined with previous studies show that homotypic chains of binding sites are important to *cis*-regulation, but their effect may vary based on the binding site and system.

A network of TFs regulates pluripotency in ES cells, and we show here that interactions between these TFs help specify their *cis*-regulatory activity. Predicting the function of genomic regulatory sequences will require not only the knowledge of which binding sites are present in a sequence, but also information about the nature of the interactions between the TFs that bind those sites. Because these interactions can be highly context dependent, analyses of synthetic regulatory elements will continue to be an important part of understanding the logic of *cis*-regulation by providing the statistical power to uncover context dependent interactions.

Methods

Cloning of plasmid libraries and overexpression plasmids

Plasmid pCF10 was constructed from pGL4.23 (Promega), first, by inserting the dsRed-Express2 gene between the Acc65I and FseI sites. Then, the *Pou5f1* basal promoter (Chr 17: 35,113,723–35,114,152; mm8) was inserted between the NcoI and HindIII sites. pCF10 served as the basic plasmid backbone for our reporter gene libraries. Array synthesized oligos (6500 unique sequences of 150 bp long) were ordered from Agilent through a limited licensing agreement. The oligos were comprised of two primer sequences, a CRE, a 9-bp barcode (BC), and multiple restriction enzyme sites (see “Synthetic CRE design” below). The OSKE library comprised of 599 CREs, each associated with 10 BCs, and the basal promoter alone associated with 30 BCs. The rest of the array contained CREs not used in this study. The array synthesized oligos were prepared as previously described (Kwasnieski et al. 2012), except using primers CF159 and CF160 with an annealing temperature of 55°C for the initial PCR step, and then purified from a polyacrylamide gel as described previously (White et al. 2013). These were cloned into plasmid pCF10 at the ApaI and SacI sites. The *Pou5f1* basal promoter and dsRed were then amplified from pCF10 using primers CF121 and CF122 and then inserted into the plasmid library from the previous step at the XbaI and HindIII sites. Plasmids without the basal promoter and dsRed were filtered out by cutting in the backbone at the SpeI site and gel extracting the band at the appropriate size. This formed the OSKE library.

The KLF4 binding sites (KBS) library was created by cloning individual CREs into the reporter plasmid. CREs with KLF4 binding sites were ordered from IDT (oligos BS300-BS308) (Supplemental Information S1). Oligonucleotides (oligos) were cloned into pCF10 at sites HindIII and ApaI, upstream of the basal promoter and dsRed gene in the same location as the CREs in the OSKE library. Two control plasmids were also constructed, with the *hsp68* promoter and the SV40 promoter. pGL-hsp68 was constructed as described previously from pCF10 (Kwasnieski et al. 2014). A plasmid with the SV40 promoter was cloned by inserting the SV40 promoter from the pbDonor-tdTomato plasmid (a gift of the Rob Mitra laboratory) using primers CF134 and CF135 at the NcoI and HindIII sites of pCF10. Oligos with BCs were then inserted into the plasmids containing the CRE inserts. First, oligos CF48 and CF49, containing random 12-bp BCs, were annealed. Next, these annealed oligos were cloned into the plasmids with CRE inserts at the XbaI and SacI sites. Twelve colonies containing random BCs for each CRE plasmid were picked and used to comprise the KBS library. The BCs in the plasmids were then Sanger sequenced, and only those plasmids with a BC insert were retained.

Overexpression constructs were created based on the pCX-OKS-2A plasmid. The individual TF genes for *Klf2*, *Klf4*, *Klf5*, and GFP were inserted between the EcoRI sites of the pCX-OKS-2A plasmid. *Klf4* sequence was taken from the pCX-OKS-2A plasmid, *Klf2* was taken from the pMXs-ms-Klf2 plasmid, and *Klf5* was taken from the pMXs-ms-Klf5 plasmid. pCX-OKS-2A (Addgene plasmid #19771), pMXs-ms-Klf2 (Addgene plasmid #50786), and pMXs-ms-Klf5 (Addgene plasmid #50787) were gifts from Shinya Yamanaka.

Synthetic CRE design

The array ordered from Agilent consisted of 150-bp oligos with the following sequence:

> ACTACAAGGGCCCA[CRE]AAGCTTCT[FILL]CGTCTAGAC
> [BC]TGAGCTCTGCAACTCCTACG

where [CRE] is the CRE comprised of concatenated building blocks of binding sites described below, [FILL] is a random filler sequence to bring the length of the sequence up to 150 bp (the filler is of variable length depending on the length of the CRE), and [BC] is a random 9-bp barcode. The CREs were chosen to fulfill every possible combination of the binding sites for the four TFs (in both orientations) of length one (eight CREs) and two binding sites (64 CREs). Additionally, a random sample of CREs with length three (191 CREs) and four (339 CREs) binding sites were chosen, enriching for homotypic combinations of binding sites. The sequence of each of the CREs is listed in Supplemental Data S1.

Each building block consisted of 20 bp with a TF binding site sequence in the middle. Each binding site, described below, consists of a 12-bp sequence. The central 10-bp sequences (underlined below) are based on binding sites from the literature or ChIP-seq data. The first position is set as a "G" in every binding site for consistency, and the last position is set as a "C" in all binding sites except for KLF4, in which it is set as a "G" to avoid a restriction site needed in the cloning.

Building block: AGCTACXXXXXXXXXXXXGT. The 12 Xs designate the binding site sequence, each of which is described below.

SOX2: GCTCATTGTTTC. Based on the canonical binding site for SOX2 (not the composite POU5F1-SOX2 site) "CATTGTT" (Chen et al. 2008b), with "CT" added before from the *Utf1* promoter and the "T" added after from the *Fgf4* promoter (Reményi et al. 2003).

POU5F1: GGGATGCTAATC. Based on the canonical binding site for POU5F1 (not the composite POU5F1-SOX2 site) "ATGCTAAT" (Chen et al. 2008b), with the "GG" added before from the *Fgf4* promoter (Reményi et al. 2003).

ESRRB: GTTCAAGGTCAC. Based on consensus binding sites "TCAAGGTC"A (van den Berg et al. 2008), with the second position ("T") from the P2 binding site in the *Pou5f1* promoter (Zhang et al. 2008).

KLF4 (OSKE library): GGGGCGGGGCCG. Based on the most common binding site matching the KLF4 PWM from KLF4 ChIP-seq peaks (Chen et al. 2008b).

KLF4 (KBS library): GGGGTGGGGCCG. Same as above, but with the fifth position changed to "T" to facilitate cloning. The fifth position can be either a "T" or a "C" according to the PWM.

Cell culture and transfection

RW4 ES cells were cultured, as described previously (Xian et al. 2005; Chen et al. 2008a), on gelatin-coated plates and in media comprised of DMEM, 10% fetal bovine serum, 10% newborn calf serum, nucleoside supplement, 1000 units/mL leukemia inhibitory factor (LIF), and 0.1 µM β-mercaptoethanol. For transfection of OSKE library, ES cells in a six-well plate were transfected using 10 µL Lipofectamine 2000 (Life Technologies), 3 µg plasmid library, and 0.3 µg CF128 (GFP plasmid control) per well. For transfection of KBS library alongside the Klf overexpression plasmids, ES cells in a six-well plate were transfected with 10 µL Lipofectamine 2000 (Life Technologies), 2.25 µg Klf overexpression plasmid (CF127, CF128, CF131, or CF136), 0.75 µg KBS library, and 0.3 µg CF128. The cells were passaged 6 h post-transfection, and RNA was extracted 26 h post-transfection using the PureLink RNA mini kit (Life Technologies). Three replicates of each sample (OSKE library and the KBS library in each overexpression condition) were transfected and processed.

Massively parallel reporter gene assays

Massively parallel reporter gene assays were used to make expression measurements of each CRE using a protocol we call CRE-seq (Kwasnieski et al. 2014). We used Illumina sequencing of both the RNA and original plasmid DNA pool. Briefly, excess DNA was removed from the RNA using the TURBO DNA-free kit (Life Technologies). cDNA was then prepared using SuperScript RT II (Life Technologies) with oligo dT primers. Both the cDNA and the plasmid DNA pool were amplified using primers CF150 and CF151b, using 21 cycles. The PCR amplification products were digested using XbaI and XhoI (NEB), and the resulting digestion products were ligated to custom Illumina adapter sequences P1_XbaI_BCX (where X is 7 through 15) and P2_XhoI, each of which is comprised of a forward (F) and reverse (R) strand that were annealed. An enrichment PCR step of 20 cycles with primers CF52 and CF53 was then used, and the resulting product was sequenced on one lane of the Illumina HiSeq for the OSKE library and on part of a lane on the Illumina MiSeq for the KBS library.

Sequencing reads were filtered to ensure that the first 13 nucleotides perfectly matched the expected sequence. For the OSKE library, this resulted in 64.3 million reads combined for the three RNA samples, and 24.5 million reads for the DNA sample. For the KBS library, this resulted in 1.79 million reads combined for the 12 RNA samples and 181,000 reads for the DNA sample. The expression of each barcode (BC) in each sample was calculated as (RNA read count)/(DNA read count) and normalized to the median expression of the BCs in each replicate. Only BCs passing a read count threshold were included for further analysis. The read count thresholds were 2000 DNA reads and three RNA reads in the OSKE library, and 100 DNA reads and five RNA reads in the KBS library. Only CREs with at least three BCs passing the read count filter in at least two replicates were included in the analysis. The overall expression of each CRE was the mean of the expression of each BC associated with it in each replicate, the standard error of the mean (SEM) was calculated from these data, and statistical comparisons between CREs were calculated from these data using a two-sided *t*-test (Supplemental Data S1). For the KBS library, there were two sets of BCs, each of which was normalized separately, and the expression of each CRE in each overexpression condition was normalized to the expression of the SV40 CRE in that condition (Supplemental Data S2).

qPCR of *Klf* genes

For quantification of gene expression level of *Klf2*, *Klf4*, and *Klf5* in overexpression conditions, RW4 cells were transfected as before but using 2.25 µg *Klf* overexpression plasmid and 0.75 µg CF128 (GFP reporter plasmid). Twenty-six hours post-transfection, the cells were resuspended in PBS and sorted on GFP using the BD FACSArias III machine into RNAprotect (Qiagen). Cells with the GFP plasmid (CF128) mimic the cells with the KBS library in the previous transfections because they were each transfected with the same amount. RNA was extracted using the RNeasy Mini Kit (Qiagen), excess DNA removed using the TURBO DNA-free kit (Life Technologies), and cDNA synthesized using SuperScript RT II (Life Technologies). qPCR was performed using Absolute SYBR Green, low ROX (Life Technologies), with primers listed in Supplemental Information S1.

Thermodynamic modeling

We used a statistical thermodynamic model based on those described previously (Buchler et al. 2003; Gertz and Cohen 2009; Gertz et al. 2009; Zeigler and Cohen 2014). The purpose of the model in this context is to provide a formal framework to model gene expression that is rooted in the biophysical principles of biomolecular interactions. The parameters of this model are proportional to the free energies of specific TF–TF interactions and allow us to infer the contribution of particular TF interactions from the observed patterns of expression in our libraries. The two key assumptions of this model are (1) that gene expression is regulated through the equilibrium binding of regulatory proteins; and (2) that the probability of RNA polymerase binding is directly proportional to the output expression. The framework and its assumptions have been discussed in detail elsewhere (Buchler et al. 2003; Bintu et al. 2005; Segal et al. 2008; Brewster et al. 2012; Sherman and Cohen 2012). The model incorporates parameters for interactions between TFs, RNAP, and binding sites on the DNA. These parameters are proportional to the free energies of interaction. For a given CRE, the statistical weight of each possible binding configuration is calculated, and the probability of RNAP being bound is the sum of the weights of all configurations in which RNAP is bound over the sum of the weights of all configurations.

The weight (W) of any given binding configuration is given as

$$W = e^{-\left(\sum q + \sum \omega\right)}, \tag{1}$$

where ω is the interaction parameter between any two proteins (either two TFs or a TF and RNAP) bound in that binding configuration and allowed to interact; and q is the interaction of a TF or RNAP and DNA. It incorporates both affinity and concentration, and is

$$q = k - \ln[\mathrm{TF}], \tag{2}$$

where k is a constant equal to $\Delta G^o/RT$. q was fixed at 0 for each TF in the reference condition (no overexpression of any TF for the OSKE library, or overexpression of GFP condition for the KBS library). The relative concentrations of the TFs in the TF overexpression conditions are used to calculate relative q values and these are fixed. The weight of the empty DNA binding state (no TFs or RNAP bound) is set as 1.

For a given CRE, the weights for all possible binding configurations are calculated, and the probability that RNAP is bound is

$$P_{\mathrm{bound}} = \frac{\sum W_{\mathrm{bound}}}{\sum W_{\mathrm{bound}} + \sum W_{\mathrm{unbound}}}, \tag{3}$$

where W_{bound} is the weight of states in which RNAP is bound; and W_{unbound} is the weight of states in which RNAP is unbound. The probability that RNAP is bound is converted to an expression measurement by scaling it so that the mean expression of all CREs in the library is the same as in the observed expression measurements.

TF interaction rules dictate when two TFs are allowed to interact, an addition to the model from previous applications (Gertz and Cohen 2009; Gertz et al. 2009; Zeigler and Cohen 2014). Only when TFs are allowed to interact does the interaction parameter contribute to the statistical weight of a given binding configuration. We have used two basic rules for TF–TF interactions based on whether two TFs can interact if another protein in bound in between. The "neighboring" interaction rule only allows TFs to interact if no other TFs are bound between them in a particular binding state. The "across" interaction rule allows TFs to interact with any other TF bound in that particular binding state (Supplemental Fig. S2). The functional consequence of the neighboring interaction rule is to impart dependence on the order of the binding sites in the CRE. The all-across interaction rule does not distinguish between different orders of binding sites for a particular combination of sites on the CRE.

Competition between KLF TFs in the KBS library was modeled by assuming that all three TFs could bind the KLF4 binding site, another addition to the model from previous uses (Gertz and Cohen 2009; Gertz et al. 2009). A further assumption for simplicity was made that all three TFs bound the site with the same affinity (somewhat supported by Jiang et al. 2008) and are present at the same concentration in the cell. Although these assumptions may be violated, in practice the model would predict the same trends regardless. The relative concentration of each TF in each overexpression condition was set based on the qRT-PCR measurements in the overexpression conditions. When modeling competition, the possible binding states include each possible KLF TF bound to each KLF4 binding site.

The model was fit with custom Python scripts using SciPy (scipy.optimize.minimize function). The parameters were fit to minimize the objective function using the L-BFGS-B and SLSQP optimization algorithms in alternating fashion until the parameter values converged. For the OSKE library, the objective function was the sum of squared errors of the expression measurements of the CREs, using the log of the observed expression and the log of the predicted expression. The initial starting values for each parameter were set to 0, but the fit was robust to different starting parameter values. Fivefold cross validation was used by splitting the data into a training set of 4/5 of the CREs and a test set of 1/5 of the CREs. Each partition of the data was used in the test set exactly once, and 95% confidence intervals for the parameter values were calculated using the asympototic normal distribution for the parameter estimate. More details can be found in Bates and Watts (1988). A set of parameter values was deemed significant if the confidence intervals for all of the parameters were significantly different than zero.

For the KBS library, the initial starting parameter values were based on an initial screen of many possible parameter values. The predicted expression of each CRE in the KBS library was calculated using a model with each of 2 million sets of random parameter values, taken from a normal distribution with mean of zero and standard deviation of 1. Each set of parameter values consisted of a value for each of the three TF–RNAP parameters and three of the six possible TF–TF interactions (three homotypic and three heterotypic), with all other parameter values set to zero. Each of the 20 possible configurations of the three TF–TF interaction parameters was sampled 100,000 times. The parameter set that predicted expression with the lowest error (using a modified objective function, see below) was used as the starting value to a fitting routine, with only those parameters with nonzero values allowed to be fit. After this fitting routine, insignificant parameters were removed from the model, and a final fitting routine was run.

For the KBS library, a modified objective function was used to take into account the relative expression in each condition as well as the overall expression. The sum of squared error of overall expression for each CRE was calculated as usual, as well as the sum of squared error of the fraction of the maximal expression in the given condition. The geometric mean of both sources of error was calculated as the objective function for fitting. This allows for better predictions of the patterns of expression.

Other statistical analysis and data sources

RNA Polymerase II ChIP-seq and DNase I hypersensitivity signal data for the mouse genome are from the ENCODE Consortium (The ENCODE Project Consortium 2012) and were downloaded from the ENCODE UCSC web portal (http://genome.ucsc.edu/ENCODE/dataMatrix/encodeDataMatrixMouse.html). TF ChIP-seq data are from Chen et al. (2008b). All genome coordinates were converted to mm9. Binding matrices were taken from JASPAR (Mathelier et al. 2014). The SOX2 binding matrix was trimmed after the eighth position to exclude the part corresponding to the POU5F1 binding site. Similarly, the POU5F1 binding matrix was trimmed before the eighth position to exclude the part corresponding to the SOX2 binding site. Permuted PWMs were created by randomly permuting the positions in the matrix (thus retaining the nucleotide content). FIMO (Grant et al. 2011) was used to find predicted binding sites using default options with a *P*-value threshold of 10^{-4}. BEDTools (Quinlan and Hall 2010) was used for manipulations and analysis of BED files. Custom scripts were used for other analysis. The sliding window approach used 200-bp windows in the genome with a 100-bp step size. For each window, we counted the number of KLF4 predicted binding sites by matches to the KLF4 PWM and the mean signals of RNA Polymerase II binding and DNase I hypersensitivity.

Data access

The sequence data from this study have been submitted to the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra/) under accession numbers SRX1684713 and SRX1686837.

Acknowledgments

We thank members of the Cohen laboratory for critical reading of the manuscript; Xuhua Chen and other members of the Mitra laboratory for cell culture assistance; and Joe Planer in the laboratory of Jeff Gordon for assistance with FACS sorting. We further thank Jamie Kwasnieski and Ilaria Mogno for help with the development of CRE-seq, and Robert Zeigler for assistance with the thermodynamic modeling code. This work was funded by National Science Foundation (NSF) Graduate Research Fellowship DGE-1143954 to C.F. and National Institutes of Health (NIH) grants R01 HG006790-01 and RO1 GM092910-02.

References

Akaike H. 1974. A new look at the statistical model identification. *IEEE Trans Automat Contr* **19:** 716–723.

Ambrosetti DC, Schöler HR, Dailey L, Basilico C. 2000. Modulation of the activity of multiple transcriptional activation domains by the DNA binding domains mediates the synergistic action of Sox2 and Oct-3 on the *fibroblast growth factor-4* enhancer. *J Biol Chem* **275:** 23387–23397.

Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339:** 1074–1077.

Arnosti DN, Kulkarni MM. 2005. Transcriptional enhancers: intelligent enhanceosomes or flexible billboards? *J Cell Biochem* **94:** 890–898.

Arnosti DN, Barolo S, Levine M, Small S. 1996. The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. *Development* **122:** 205–214.

Bates DM, Watts DG. 1988. *Nonlinear regression analysis and its applications*. John Wiley & Sons, New York.

Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, Phillips R. 2005. Transcriptional regulation by the numbers: models. *Curr Opin Genet Dev* **15:** 116–124.

Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG, et al. 2005. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122:** 947–956.

Brewster RC, Jones DL, Phillips R. 2012. Tuning promoter strength through RNA polymerase binding site design in *Escherichia coli*. *PLoS Comput Biol* **8:** e1002811.

Brewster RC, Weinert FM, Garcia HG, Song D, Rydenfelt M, Phillips R. 2014. The transcription factor titration effect dictates level of gene expression. *Cell* **156:** 1312–1323.

Buchler NE, Gerland U, Hwa T. 2003. On schemes of combinatorial transcription logic. *Proc Natl Acad Sci* **100:** 5136–5141.

Chen CT, Gottlieb DI, Cohen BA. 2008a. Ultraconserved elements in the *Olig2* promoter. *PLoS One* **3:** e3946.

Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, et al. 2008b. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133:** 1106–1117.

Chew JL, Loh YH, Zhang W, Chen X, Tam WL, Yeap LS, Li P, Ang YS, Lim B, Robson P, et al. 2005. Reciprocal transcriptional regulation of *Pou5f1* and *Sox2* via the Oct4/Sox2 complex in embryonic stem cells. *Mol Cell Biol* **25:** 6031–6046.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74.

Erceg J, Saunders TE, Girardot C, Devos DP, Hufnagel L, Furlong EE. 2014. Subtle changes in motif positioning cause tissue-specific effects on robustness of an enhancer's activity. *PLoS Genet* **10:** e1004060.

Gertz J, Cohen BA. 2009. Environment-specific combinatorial *cis*-regulation in synthetic promoters. *Mol Syst Biol* **5:** 244.

Gertz J, Siggia ED, Cohen BA. 2009. Analysis of combinatorial *cis*-regulation in synthetic and genomic promoters. *Nature* **457:** 215–218.

Goldwater MB, Markman AB, Stilwell CH. 2010. The empirical case for role-governed categories. *Dev Cell* **18:** 359–376.

Gotea V, Visel A, Westlund JM, Nobrega MA, Pennacchio LA, Ovcharenko I. 2010. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res* **20:** 565–577.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27:** 1017–1018.

He X, Samee MA, Blatti C, Sinha S. 2010. Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. *PLoS Comput Biol* **6:** e1000935.

Hutchins AP, Choo SH, Mistri TK, Rahmani M, Woon CT, Ng CKL, Jauch R, Robson P. 2013. Co-motif discovery identifies an esrrb-Sox2-DNA ternary complex as a mediator of transcriptional differences between mouse embryonic and epiblast stem cells. *Stem Cells* **31:** 269–281.

Ivanova N, Dobrin R, Lu R, Kotenko I, Levorse J, DeCoste C, Schafer X, Lun Y, Lemischka IR. 2006. Dissecting self-renewal in stem cells with RNA interference. *Nature* **442:** 533–538.

Jiang J, Chan YS, Loh YH, Cai J, Tong GQ, Lim CA, Robson P, Zhong S, Ng HH. 2008. A core Klf circuitry regulates self-renewal of embryonic stem cells. *Nat Cell Biol* **10:** 353–360.

Kim TK, Maniatis T. 1997. The mechanism of transcriptional synergy of an in vitro assembled interferon-β enhanceosome. *Mol Cell* **1:** 119–129.

Kim J, Chu J, Shen X, Wang J, Orkin SH. 2008. An extended transcriptional network for pluripotency of embryonic stem cells. *Cell* **132:** 1049–1061.

Kinney JB, Murugan A, Callan CG, Cox EC. 2010. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc Natl Acad Sci* **107:** 9158–9163.

Kulkarni MM, Arnosti DN. 2003. Information display by transcriptional enhancers. *Development* **130:** 6569–6575.

Kuroda T, Tada M, Kubota H, Kimura H, Hatano S, Suemori H, Nakatsuji N, Tada T. 2005. Octamer and Sox elements are required for transcriptional *cis* regulation of *Nanog* gene expression. *Mol Cell Biol* **25:** 2475–2485.

Kwasnieski JC, Mogno I, Myers CA, Corbo JC, Cohen BA. 2012. Complex effects of nucleotide variants in a mammalian *cis*-regulatory element. *Proc Natl Acad Sci* **109:** 19498–19503.

Kwasnieski JC, Fiore C, Chaudhari HG, Cohen BA. 2014. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res* **24:** 1595–1602.

Levo M, Zalckvar E, Sharon E, Dantas Machado AC, Kalma Y, Lotam-Pompan M, Weinberger A, Yakhini Z, Rohs R, Segal E. 2015. Unraveling determinants of transcription factor binding outside the core binding site. *Genome Res* **25:** 1018–1029.

Liu F, Posakony JW. 2012. Role of architecture in the function and specificity of two notch-regulated transcriptional enhancer modules. *PLoS Genet* **8:** e1002796.

Loh YH, Wu Q, Chew JL, Vega VB, 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.

Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen CY, Chou A, Ienasescu H, et al. 2014. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res* **42:** D142–D147.

Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG, Kinney JB, et al. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30:** 271–277.

Mogno I, Kwasnieski JC, Cohen BA. 2013. Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Res* **23:** 1908–1915.

Nakatake Y, Fukui N, Iwamatsu Y, Masui S, Takahashi K, Yagi R, Yagi K, Miyazaki JI, Matoba R, Ko MSH, et al. 2006. Klf4 cooperates with Oct3/4 and Sox2 to activate the *Lefty1* core promoter in embryonic stem cells. *Mol Cell Biol* **26:** 7772–7782.

Panne D, Maniatis T, Harrison SC. 2007. An atomic model of the interferon-β enhanceosome. *Cell* **129:** 1111–1123.

Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee SI, Cooper GM, et al. 2012. Massively parallel functional dissection of mammalian enhancers *in vivo*. *Nat Biotechnol* **30:** 265–270.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842.

Reményi A, Lins K, Nissen LJ, Reinbold R, Schöler HR, Wilmanns M. 2003. Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. *Genes Dev* **17:** 2048–2059.

Rodda DJ, Chew JL, Lim LH, Loh YH, Wang B, Ng HH, Robson P. 2005. Transcriptional regulation of *Nanog* by OCT4 and SOX2. *J Biol Chem* **280:** 24731–24737.

Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U. 2008. Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature* **451:** 535–540.

Senger K, Armstrong GW, Rowell WJ, Kwan JM, Markstein M, Levine M. 2004. Immunity regulatory DNAs share common organizational features in *Drosophila*. *Mol Cell* **13:** 19–32.

Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, Keren L, Yakhini Z, Weinberger A, Segal E. 2012. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol* **30:** 521–530.

Shea MA, Ackers GK. 1985. The O_R control system of bacteriophage λ. A physical-chemical model for gene regulation. *J Mol Biol* **181:** 211–230.

Sherman MS, Cohen BA. 2012. Thermodynamic state ensemble models of *cis*-regulation. *PLoS Comput Biol* **8:** e1002407.

Smith RP, Taher L, Patwardhan RP, Kim MJ, Inoue F, Shendure J, Ovcharenko I, Ahituv N. 2013. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat Genet* **45:** 1021–1028.

Thanos D, Maniatis T. 1995. Virus induction of human IFNβ gene expression requires the assembly of an enhanceosome. *Cell* **83:** 1091–1100.

van den Berg DL, Zhang W, Yates A, Engelen E, Takacs K, Bezstarosti K, Demmers J, Chambers I, Poot RA. 2008. Estrogen-related receptor β interacts with Oct4 to positively regulate *Nanog* gene expression. *Mol Cell Biol* **28:** 5986–5995.

Wei Z, Yang Y, Zhang P, Andrianakos R, Hasegawa K, Lyu J, Chen X, Bai G, Liu C, Pera M, et al. 2009. Klf4 interacts directly with Oct4 and Sox2 to promote reprogramming. *Stem Cells* **27:** 2969–2978.

White MA, Myers CA, Corbo JC, Cohen BA. 2013. Massively parallel in vivo enhancer assay reveals that highly local features determine the *cis*-regulatory function of ChIP-seq peaks. *Proc Natl Acad Sci* **110:** 11952–11957.

Xian HQ, Werth K, Gottlieb DI. 2005. Promoter analysis in ES cell-derived neural cells. *Biochem Biophys Res Commun* **327:** 155–162.

Yáñez-Cuna JO, Dinh HQ, Kvon EZ, Shlyueva D, Stark A. 2012. Uncovering *cis*-regulatory sequence requirements for context-specific transcription factor binding. *Genome Res* **22:** 2018–2030.

Zeigler RD, Cohen BA. 2014. Discrimination between thermodynamic models of *cis*-regulation using transcription factor occupancy data. *Nucleic Acids Res* **42:** 2224–2234.

Zhang X, Zhang J, Wang T, Esteban MA, Pei D. 2008. *Esrrb* activates *Oct4* transcription and sustains self-renewal and pluripotency in embryonic stem cells. *J Biol Chem* **283:** 35825–35833.

Received October 12, 2015; accepted in revised form April 13, 2016.