Jin_Supplementary Material

The human promoter array platform. A set of 24,134 promoters with an average length of 1500 bp for each promoter from human genome build 35 (UCSC HG17 Assembly), corresponding to 18,855 human genes, was used to construct an array platform by NimbleGen Systems Inc. Each promoter included 15 50-mer probes and random 50-mer probes were also included as non-specific hybridization controls, therefore a total of ~380,000 probes were spotted on the array. The measurements of the intensity of the hybridized arrays were performed by scanning on an Axon GenePix 4000B scanner (Axon Instruments Inc.) at wavelengths of 532 nm for control (Cy3), and 635 nm (Cy5) for each experimental sample. Data points were extracted from the scanned images using the NimbleScan 2.0 program (NimbleGen Systems, Inc.). Each pair of N probe signals was normalized by converting into a scaled log ratio using the following formula: $S_i = Log_2 (Cy5I(i) /Cy3(i))$

Strategy for selecting datasets for training ChIPModules. The identification of target genes for a specific transcription factor is partly dependent on the array platform. Many bioinformatics programs such as HMM (Li et al. 2005), MAT (Johnson et al. 2006), Mpeaks (Kim et al. 2005), and Tamalpais Peaks (Bieda et al. 2006) have been developed to identify binding sequences for high density tiling arrays. For spotted array platforms, approaches such as a median percentile rank (Lieb et al. 2001), a single array-error model (Ren et al. 2000); and a sliding window analysis (ChIPOTle: (Buck and Lieb 2004; Buck et al. 2005)) have proven to be effective in identifying the target genes. In our previous studies (Jin et al. 2006; Squazzo et al. 2006), we developed two approaches to identify target genes using promoter arrays. We employed both of these approaches in this study to select the training set and to identify the targets genes for OCT4 and SRY. For training datasets, we used two biological replicates (replicates A and B) of OCT4 ChIP-chip data. One way to examine reproducibility is to compare the two datasets. All promoters on an array were ranked by enrichment (using the average intensity of all 15 probes/promoter) and then the two ranked lists were compared using intervals of 100 (**Supplementary Figure 1)**. The results show that an average of 58.5% of the Top 500 (293 genes) were the same in the two datasets. Another way to

compare arrays is to use a peaksCalling programs that we developed in a previous study (Jin et al. 2006). Using this program, we identified a set of 1165 peaks at the Top 5% confidence level. Of the 293 promoters identified in the Top 500 of both Replicates A and B, 288 (98.3%) were also among the Top 5% level of peaks picked by our peaksCalling program. Based on the facts that these 293 targets have both strong average intensities and high confidence level of binding signals, we chose them as a initial training data (**Dataset 3**) for our **ChIPModules** approach.

Strategy for identifying common genes bound by both OCT4 and SRY. We applied a similar strategy to the average ranking method described above to identify the commonly bound target promoters for OCT4 and SRY. We performed two ChIP-chip arrays (using two biologically independent ChIP assays) for both OCT4 and SRY. Each of the duplicate samples was from cells that were grown, cross-linked, and assayed independently. We found that 1104 (55.2%) were overlapped at the Top 2000 ranked OCT4 targets in both replicates, while 1344 (67.2%) were overlapped at the Top 2000 of SRY targets respectively. Of 1104 OCT4 targets, 538 (49%) were also SRY targets. The results for promoters bound by OCT4, SRY and both factors are shown in **Supplementary Table S1**.

Whole Genome Amplification Protocol for ChIP-chip

A. Library Preparation

1. Add 2 ul 1X Library Preparation Buffer to 10 ul of input material [For the "input" sample, measure the concentration of reverse crosslinked, QIAquick purified DNA and add 10 ng to a total volume of 10 ul with H2O. For the ChIP sample, the concentration of nucleic acid is usually too low to get an accurate quantitation. Typically the entire 50ul of reverse crosslinked, QIAquick purified DNA is lyophilized and resuspended in 10 ul of H20]

Transfer samples to strip tubes or individual thin walled 0.2 ml PCR tubes

2. Add 1 ul Library Stabilization Solution, vortex or mix by pipetting. Quick spin and place at 95° for 2 minutes in thermal cycler
3. Immediately cool on ice, quick spin again
4. Add 1 ul Library Preparation Enzyme, vortex or mix by pipetting and quick spin if necessary
5. Incubate in thermal cycler as follows:
    - 16° for 20' (cycler should be precooled to this temperature)
    - 24° for 20'
    - 37° for 20'
    - 75° for 5'
    - 4° hold

6. Quick spin if necessary and either proceed to first amplification or freeze at -20° for up to three days

B. Amplification (round 1)

7. Prepare master mix for each sample containing:
7.5 ul of 10X Amplification Master Mix
47.5 ul Nuclease-free H20
5 ul WGA DNA polymerase
From O'Geen et al., BioTechniques 41(5), (November 2006)
[For multiple samples, multiply above volumes by the number of samples then add 1/10 volume extra of each component]
8. Add 60 ul master mix to each sample, vortex or mix by pipetting and quick spin if necessary
9. Incubate in thermal cycler as follows:
    95° for 3' , then 14 cycles of
    94° for 15"
    65° for 5', then
    4° hold

At this point, amplified material is stable and can be stored at –20 ° indefinitely
10. Purify samples using QIAquick PCR cleanup columns or analogous product. It is important to elute the samples in water so that the subsequent labeling reactions are efficient.
[Since the amplified material contains both single- and double-stranded DNA that can be effectively labeled, the column purification method used should recover both.]
[At this stage, the purification column eluates for total and immunoprecipitated samples should be readily quantifiable by nanodrop, spectrometer, or dye intercalation, eg, picogreen (dye intercalation may underestimate amount due to single strand product).
Optimally, total recovery for immunoprecipitated samples will be in the 1-4 ug range.
This gives enough material for several labelings for downstream microarray analysis. If yields are less, or more product is desired, re-amplify material using Sigma GenomePlex WGA Reamplification Kit]

C. Reamplification (round 2)

1. Add 15 ng purified amplification product in 10 ul volume to strip tubes or individual thin walled 0.2 ml PCR tubes
[For input material start with the high concentration primary amplified stock]
2. Prepare master mix for each sample containing:
    7.5 ul of 10X Amplification Master Mix
    47.5 ul Nuclease-free H20
    5 ul WGA DNA polymerase

For multiple samples, multiply above volumes by the number of samples then add 1/10 volume extra of each component
3. Add 60 ul master mix, vortex or mix by pipetting and quick spin if necessary
From O'Geen et al., BioTechniques 41(5), (November 2006)
4. Incubate in thermal cycler as follows:
    95° for 3' , then 14 cycles of
    94° for 15"
    65° for 5', then
    4° hold

At this point, amplified material is stable and can be stored at –20 ° indefinitely
5. Purify samples using QIAquick PCR cleanup columns or analogous product.
[Since the amplified material contains both single- and double-stranded DNA that can be effectively labeled, the column purification method used should recover both.]

Supplementary References

Bieda, M., X. Xu, M. Singer, R. Green, and P.J. Farnham. 2006. Unbiased location analysis of E2F1 binding sites suggests a widespread role for E2F1 in the human genome. *Genome Research* **16:** 595-605.

Buck, M.J. and J.D. Lieb. 2004. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* **83:** 349-360.

Buck, M.J., A.B. Nobel, and J.D. Lieb. 2005. ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. *Genome Biol.* **6:** R97.

Jin, V.X., A. Rabinvich, S.L. Squazzo, R. Green, and P.J. Farnham. 2006. A computational genomics approach to identify cis-regulatory modules from chromatin immunoprecipitation microarray data-a case study using E2F1. *Genome Research* Oct 19 (In Advance).

Johnson, W.E., W. Li, C.A. Meyer, R. Gottardo, J.S. Carroll, M. Brown, and X.S. Liu. 2006. Model-based analysis of tiling-arrays for ChIP-chip. *Proc Natl Acad Sci U S A.* **103:** 12457-12462.

Kim, T.H., L.O. Barrera, M. Zheng, C. Qu, M.A. Singer, T.A. Richmond, Y. Wu, R.D. Green, and B. Ren. 2005. A high-resolution map of active promoters in the human genome. *Nature* **436:** 876-880.

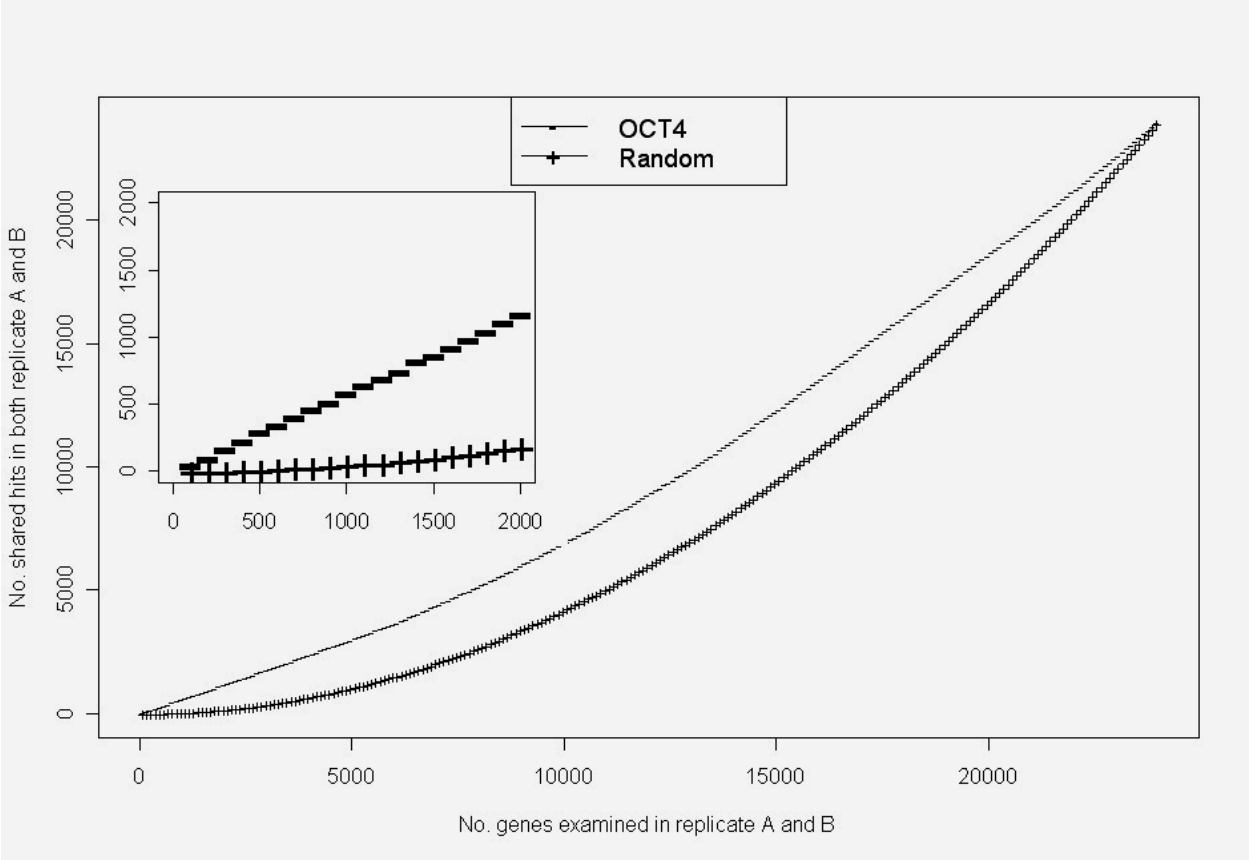
Li, W., C.A. Meyer, and X.S. Liu. 2005. A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics* **21 Suppl 1:** i274-282.

Lieb, J.D., X. Liu, D. Botstein, and P.O. Brown. 2001. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat Genet.* **28:** 327-334.

O'Geen, H., C.M. Nicolet, K. Blahnik, R. Green, and P.J. Farnham. 2006. Improvements in sample preparation for ChIP-chip assays. *BioTechniques***: In press**.

Ren, B., F. Robert, J.J. Wyrick, O. Aparicio, E.G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T.L. Volkert, C.J. Wilson, S.P. Bell, and R.A. Young. 2000. Genome-wide location and function of DNA binding proteins. *Science* **290:** 2306– 2309.

Squazzo, S.L., V.M. Komashko, H. O'Geen, S. Krig, V.X. Jin, S.-W. Jang, R. Green, R. Margueron, D. Reinberg, and P.J. Farnham. 2006. Suz12 silences large regions of the genome in a cell type-specific manner. *Genome Research* **16:** 890-900.



Jin_Supplementary Figure 1: The number of genes in common when comparing ranked (based on the average intensity of all 15 probes for each promoter) lists for two biological replicates of OCT ChIP-chip data ("-" line) and the number of genes in common for randomly simulated datasets ("+" line) are shown.

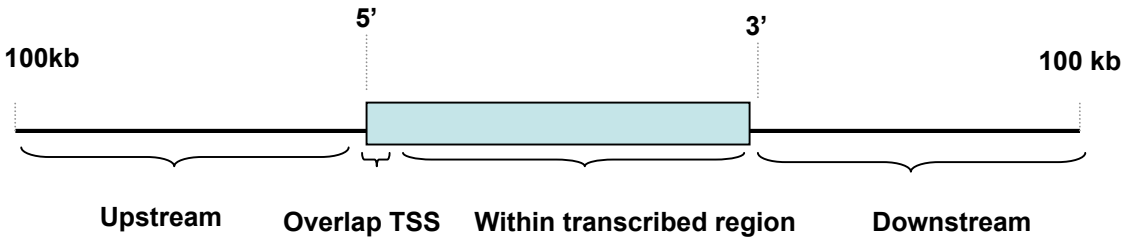
NAME OCT_Q6
MATR_LENGTH 11
CORE_START 6; CORE_LENGTH 5
MAXIMAL 22599.513672
MINIMAL 508.774689
THRESHOLD 0.703978
WEIGHTS
1 A:115.647272 C:80.953090 G:115.647272 T:323.812362
2 A:3.233523 C:2.771591 G:3.464489 T:3.233523
3 A:1175.785992 C:222.445998 G:63.556000 T:286.001998
4 A:385.094546 C:128.364849 G:0.000000 T:3016.573945
5 A:316.677916 C:226.198511 G:45.239702 T:1900.067493
6 A:320.009346 C:106.669782 G:106.669782 T:2400.070095
7 A:274.590721 C:74.888379 G:848.734956 T:174.739550
8 A:220.835217 C:3680.586943 G:0.000000 T:147.223478
9 A:4702.786356 C:0.000000 G:0.000000 T:177.463636
10 A:80.407346 C:80.407346 G:80.407346 T:4181.182016
11 A:366.449601 C:54.288830 G:135.722074 T:190.010904

NAME OCT-C
MATR_LENGTH 13
CORE_START 5; CORE_LENGTH 5
MAXIMAL 7553.789551
MINIMAL 34.912647
THRESHOLD 0.565636
WEIGHTS
1 A:28.330864 C:68.335092 G:12.722656 T:24.002537
2 A:50.954275 C:0.000000 G:49.176800 T:502.432850
3 A:13.249834 C:34.197896 G:15.988627 T:11.769406
4 A:622.645219 C:99.509510 G:0.000000 T:0.000000
5 A:0.000000 C:0.000000 G:0.000000 T:1016.999994
6 A:0.000000 C:0.000000 G:0.000000 T:1016.999994
7 A:105.727082 C:0.000000 G:0.000000 T:600.974995
8 A:0.000000 C:0.000000 G:600.974995 T:105.727082
9 A:0.000000 C:1016.999994 G:0.000000 T:0.000000
10 A:719.213498 C:0.000000 G:72.466799 T:0.000000
11 A:0.000000 C:0.000000 G:0.000000 T:1016.999994
12 A:260.383648 C:0.000000 G:98.228562 T:37.810200
13 A:21.482436 C:76.631375 G:10.420585 T:54.347356

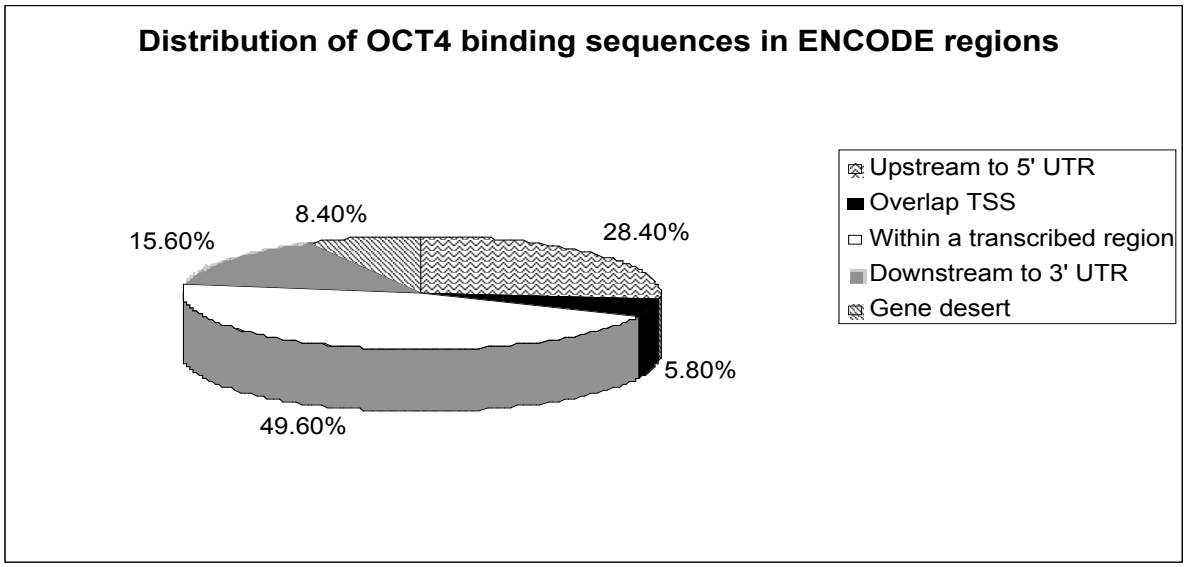
Jin_Supplementary Figure 2: **Positional Weight Matrices (PWMs) for OCT4**

Shown are the two OCT related PWMs in the TRANSFAC database, OCT_Q6 and OCT_C. We used OCT_Q6 to predict OCT4 binding sites in the promoter sequences. However, we obtained similar results if we used the other PWM (data not shown).

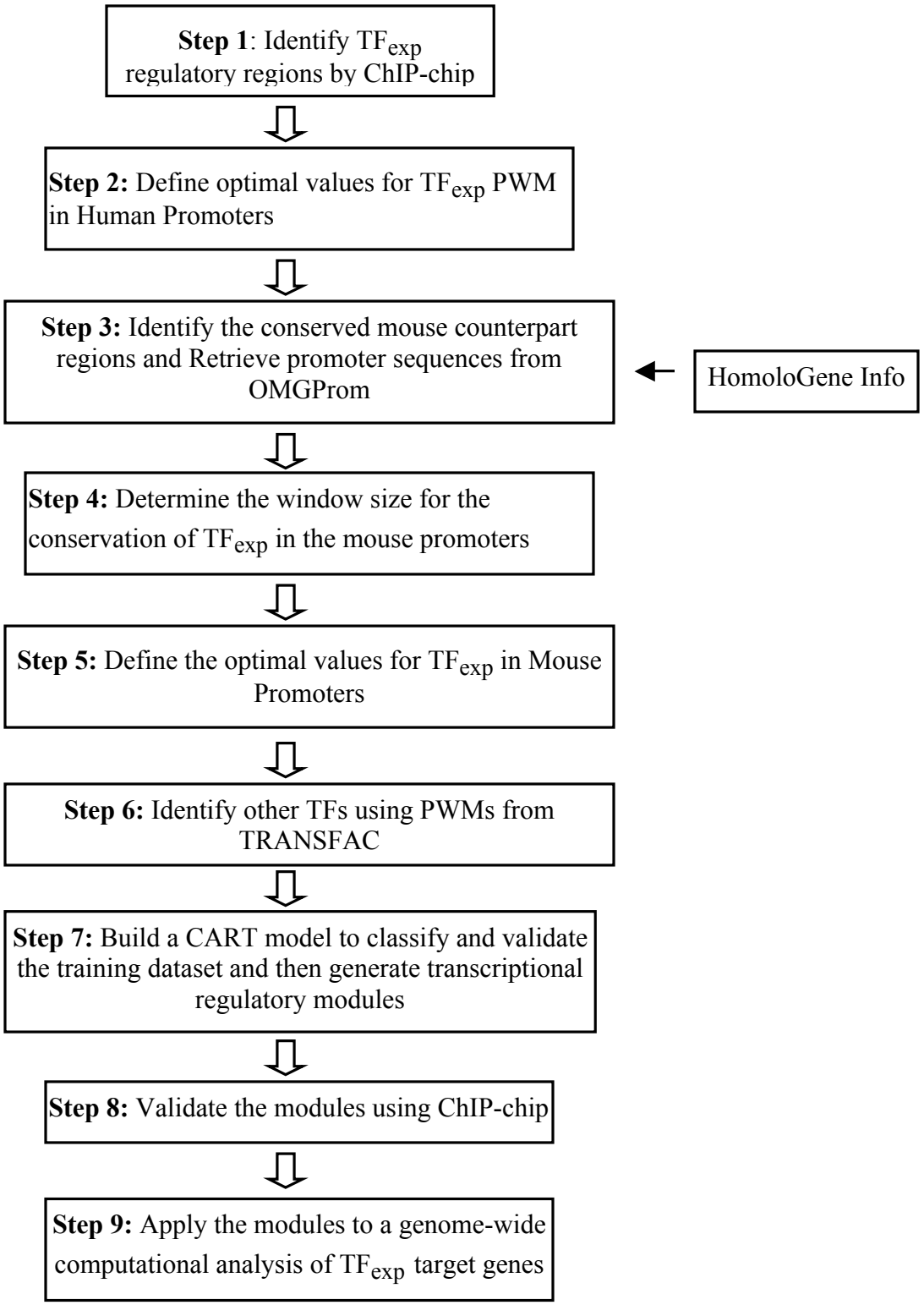
Jin_ Supplementary Figure 3A



Jin_ Supplementary Figure 3B



Distribution of OCT4 binding sites on the basis of gene structure. **A)** A schematic identifying the different categories is shown; the categories include upstream regions up to 100 kb 5' of a gene, sites overlapping with transcription start sites, sites within a transcribed region, and sites within 100 kb downstream of a gene. **B)** A pie chart is shown which indicates the percentage of OCT4 binding sites in the different categories described in panel A. Also shown in the percentage of OCT4 binding sites in gene deserts (defined as a site that is farther than 100 kb from either a 5' or 3' end of a gene). The OCT4 binding regions were defined as peaks detected on duplicate ENCODE arrays using a peak calling program developed for ChIP-chip experiments (Bieda et al., 2006); the gene list was based on Gencode Genes (Harrow et al., 2006).


Step 1: Identify TFexp regulatory regions by ChIP-chip
Step 2: Define optimal values for TFexp PWM in Human Promoters
Step 3: Identify the conserved mouse counterpart regions and Retrieve promoter sequences from OMGProm
HomoloGene Info
Step 4: Determine the window size for the conservation of TFexp in the mouse promoters
Step 5: Define the optimal values for TFexp in Mouse Promoters
Step 6: Identify other TFs using PWMs from TRANSFAC
Step 7: Build a CART model to classify and validate the training dataset and then generate transcriptional regulatory modules
Step 8: Validate the modules using ChIP-chip
Step 9: Apply the modules to a genome-wide computational analysis of TFexp target genes

Jin_Supplementary Figure 4: **Flowchart of the ChIPModules approach.** Because we employed our **ChIPModules** approach from a previous study in this current study, we have summarized the **ChIPModules** approach here; details of the approach are in Jin et al. (Jin et al. 2006). The approach begins with a set of experimentally defined binding sites (TF_{exp}), refines the set to include only those sites conserved in the orthologous mouse promoters, searches for nearby binding sites for other factors, builds a CART model to generate a high confidence set of co-occurring binding sites, validates the co-localization of the factors using additional ChIP-chip assays, and then searches for the validated ChIPModules in a large promoter database.

Step 1: The first step is to identify the TF_{exp} regulatory regions using ChIP-chip data. It is important to choose a positive control training set based on either a low p value or a high enrichment value (depending on the analysis program used to identify target genes). In our case, we use the E2F1 binding sites identified in a previous study to have a p-value less than 0.0001 (Bieda et al. 06). The negative training set should be from unenriched promoters from the same ChIP-chip experiment.

Step 2: Identify TF_{exp} binding sites by using the TF_{exp} PWM either constructed by yourself or from TRANSFAC; Define the optimal values for a match to the core consensus and PWM for the TF_{exp}. For this step, it is important that the scores chosen should identify a high percentage of the positive training set and a relatively lower percentage of the negative training set.

Step 3: Identify the conserved mouse counterpart promoters. First, use HomoloGene Information to identify the appropriate mouse gene for the human target promoters. Then, retrieve the human and mouse promoter regions from OMGProm.

Step 4: Determine the window size for the conservation of the TF_{exp} in the mouse promoters. For this step, the window size chosen should identify sites in a high percentage of the positive training set and in a relatively lower percentage of the negative training set.

Step 5: Define the optimal values for the match to the core consensus and PWM for the TF_{exp} in the mouse promoters. As in step 2, it is important that the scores chosen should identify a high percentage of the positive training set and a relatively lower percentage of the negative training set.

Step 6: Identify other TFs using PWMs from TRANSFAC database (http://www.gene-regulation.com/pub/databases.html#transfac). Select those TFs within a distance of the TF_{exp}. The value of the distance can be chosen varying from 220 bp to 500 bp in this step.

Step 7: Build a CART model to classify the training dataset for those various values of the distance. Determine the value of the distance for those TFs nearby the TF_{exp} (**Figure 3**) by maximizing the sensitivity and specificity calculated from the CART. Validate the training dataset by ROC method using a range of separation values. Then generate transcriptional regulatory modules using CART.

Step 8: Validate the modules using ChIP-chip and antibodies to the TF_{exp} and the newly identified co-localizing TF.

Step 9: Apply the modules to a genome-wide computational analysis of TF_{exp} target genes.