Supplementary Methods and Figures

Measurement of Intensity of the Array Data

The hybridized arrays were scanned 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:

$S_i = Log_2 (Cy5I(i) /Cy3(i))$

Algorithms for peaksPicking program

Peaks are usually bell-shaped, and also have neighboring effects. A plot of the normalized data points on the arrays (from the GFF file produced by NimbleScan 2.0 program) shows a normal distribution using one-sample kolmogorov-smirnov good fitness test ($p < 10^{-13}$) shown in **Supplementary Figure 1**.

Algorithm:

1. Defining a confidence interval:
    a. Sort the value from an array data S_0 ($S=\{S_{01}, S_{02}, \ldots S_{0n}\}$) of input GFF file
    b. Retrieve the value (S_d) of that confidence interval.
2. Picking peaks:
    a. Average the values of k (k={1,2,3,4…}) consecutive oligos, forming a new array data S1.
    b. Calculate each mean value $\bar{S}$ from new array data S1 ($S1=\{S_{11}, S_{12}, \ldots S_{1n}\}$) with a minimum of 5 consecutive oligos for each promoter.
    c. If "$\bar{S} > S_d$" continue adding a new oligo, repeat step b.
    else Output the coordinates of these oligos and values.

d. Continue step b until the end of array data S1.

3. The computation formula for a mean $\bar{S}$ value for a peak picked:

$$\bar{S} = \frac{\sum_{i=1}^{k-1} iSi + \sum_{i=k}^{m} kSi + \sum_{i=m+1}^{m+k-1} (m-i+k)Si}{km}$$

Where k is number of consecutive oligos, m is the number of oligos picked as peak regions, Si is the value of an oligo. $\sum_{i=1}^{k-1} iSi$ = 0 and $\sum_{i=m+1}^{m+k-1} (m-i+k)Si$ = 0 when k=1.

4. p value for the peaks:

For a mean value $\bar{S}$ of a peak, the Z score is defined in the following:

$$Z = \left(\bar{S} - \mu \right) / \sigma$$

where μ and σ is the mean value and standard deviation for the whole data points. The corresponding p value can be calculated using Z score by a normal distribution by one-tailed side.

At the Top 2%, 5%, 10% and 20% level of peaks, the p values are less than 0.02, 0.05. 0.1 and 0.2 respectively.

Implementation:

The program peaksPicking was implemented in Perl language. It is run on both Windows and Linux/Unix platforms. The input file is a GFF file from NimbleGen Arrays. The peaks detected can be viewed by the visualization software Signalmap developed by Nimblegen Inc.

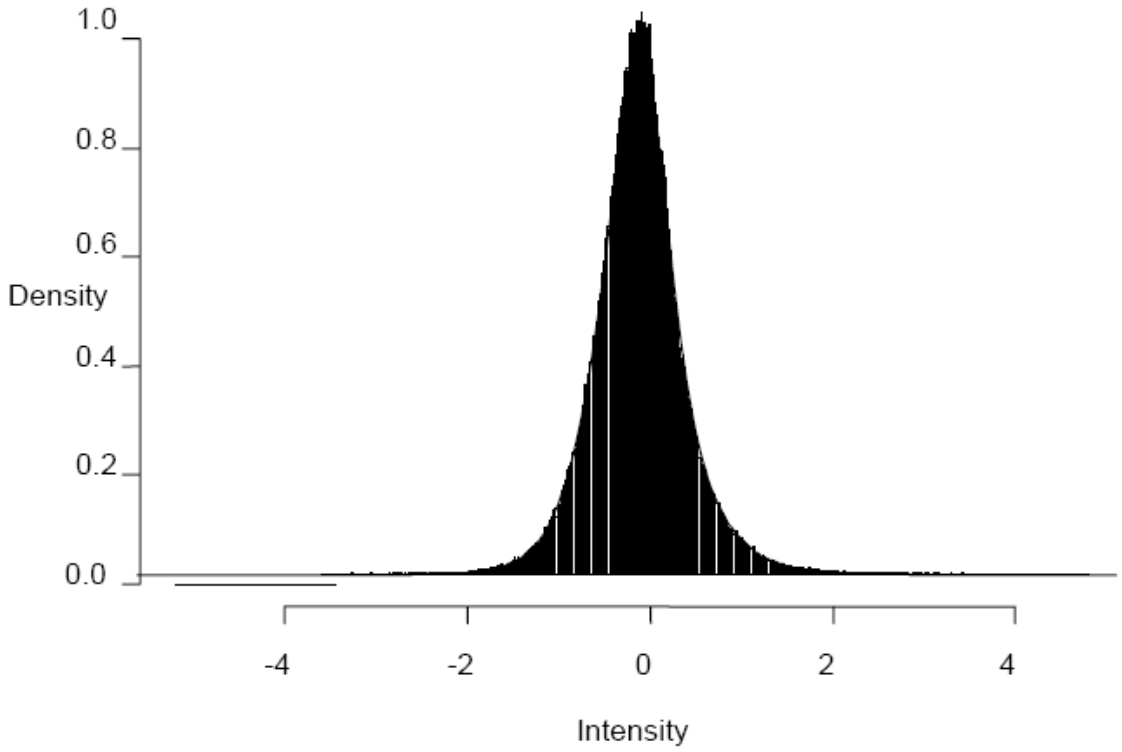
Strategy for common genes bound by both E2F1 and AP-2α:

We have performed three ChIP-chip arrays (using two biologically independent ChIP assays, one of which was run in duplicate as technical replicate) for both E2F1 and AP-2α. Each of the three arrays used a ChIP sample from cells that

were grown, cross-linked, and assayed independently. We first applied the program peaksPicking to each array data to identify the peaks; Secondly, we found common promoters with peaks from at least two of three arrays for E2F1 and AP-2α respectively; Finally we chose the overlap peaks of E2F1 and AP-2α for each promoter using three distance apart (270 bp, 500 bp and 1000 bp) for overlap peaks. The results for promoters with E2F1, AP-2α and common promoters are shown in **Table 3 and Supplementary Table S1**. A capture of peaks picked for overlap of E2F1 and AP-2α at the Top 10% level is shown in **Supplementary Figure 2**.

Positional Weight Matrices (PWMs) for E2F1

There are three E2F1 related PWMs in the TRANSFAC database, E2F1_Q3, E2F1_Q4 and E2F1_Q6. We used E2F1_Q3 to predict E2F1 binding sites for each promoter sequence. The similar results were obtained for other two PWMs (data not shown). All three E2F1 PWMs in the TRANSFAC database are shown in **Supplementary Figure 3**.



Supplementary Figure 1: Intensity of Data Points vs Density showing the normal distribution of the data.

Supplementary Figure 2: A capture of overlap peaks of E2F1 and AP-2α picked at the Top 10% level for promoters LENG9 and GATA1.

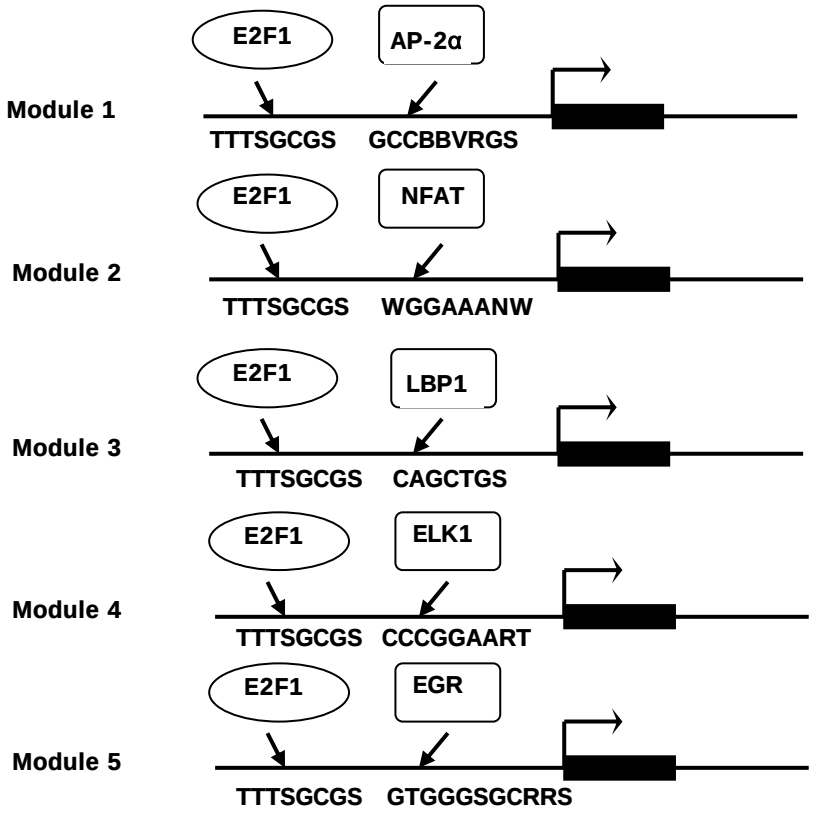
NAME E2F1_Q3
MATR_LENGTH 8
CORE_START 3
CORE_LENGTH 5
MAXIMAL 3341.360840
MINIMAL 8.687737
THRESHOLD 0.824261
WEIGHTS
1 A:8.687737 C:34.750948 G:26.063211 T:43.438685
2 A:0.000000 C:35.213110 G:176.065550 T:246.491770
3 A:138.061781 C:0.000000 G:0.000000 T:759.339793
4 A:0.000000 C:351.495392 G:301.281764 T:0.000000
5 A:0.000000 C:259.690849 G:415.505358 T:0.000000
6 A:0.000000 C:610.325081 G:183.097524 T:0.000000
7 A:0.000000 C:183.097524 G:610.325081 T:0.000000
8 A:0.000000 C:304.439497 G:152.219749 T:38.054937

NAME E2F1_Q4
MATR_LENGTH 8
CORE_START 2
CORE_LENGTH 5
MAXIMAL 2673.393311
MINIMAL 3.903595
THRESHOLD 0.715482
WEIGHTS
1 A:3.903595 C:3.903595 G:3.903595 T:7.807191
2 A:0.000000 C:0.000000 G:0.000000 T:500.000000
3 A:0.000000 C:0.000000 G:0.000000 T:500.000000
4 A:0.000000 C:102.904941 G:154.357411 T:0.000000
5 A:0.000000 C:63.903595 G:255.614381 T:0.000000
6 A:0.000000 C:500.000000 G:0.000000 T:0.000000
7 A:0.000000 C:0.000000 G:500.000000 T:0.000000
8 A:0.000000 C:63.903595 G:255.614381 T:0.000000

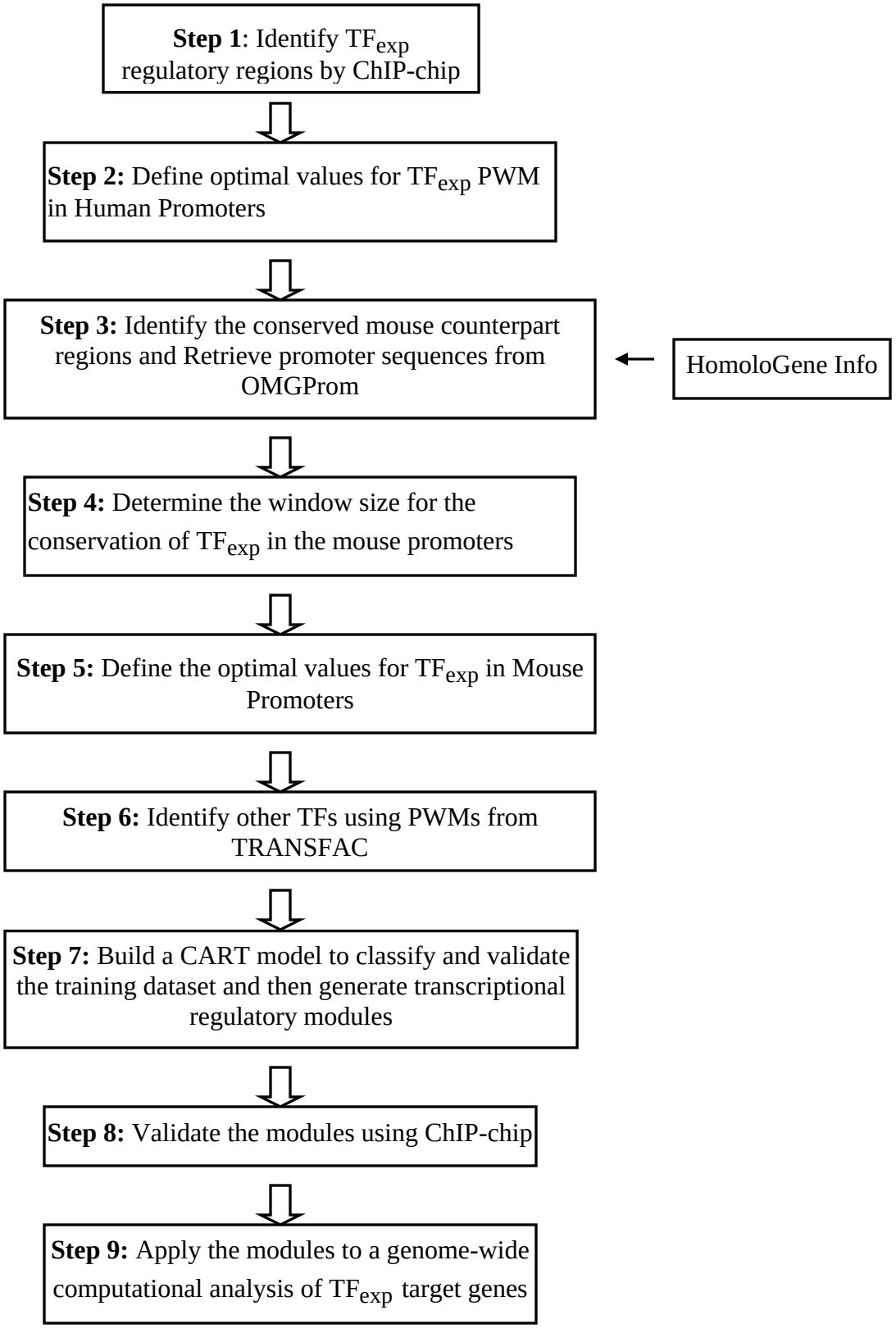
NAME E2F1_Q6
MATR_LENGTH 8
CORE_START 3
CORE_LENGTH 5
MAXIMAL 9146.661133
MINIMAL 36.041374
THRESHOLD 0.788279
WEIGHTS
1 A:36.041375 C:108.124126 G:72.082751 T:270.310316

2 A:0.000000 C:95.060147 G:237.650369 T:950.601475
3 A:224.511250 C:0.000000 G:0.000000 T:1796.089998
4 A:0.000000 C:819.904748 G:563.684514 T:0.000000
5 A:0.000000 C:486.766875 G:973.533749 T:0.000000
6 A:0.000000 C:1796.089998 G:224.511250 T:0.000000
7 A:0.000000 C:224.511250 G:1796.089998 T:0.000000
8 A:0.000000 C:744.040598 G:393.903846 T:43.767094

Supplementary Figure 3: Three E2F1 PWMs: E2F1_Q3, E2F1_Q4 and E2F1_Q6.



Supplementary Figure 4: Five *cis*-regulatory modules identified from the ChIPModules approach from E2F1 ChIP-chip in both HeLa and MCF-7 cells.


Step 1: Identify TF_{exp} regulatory regions by ChIP-chip
Step 2: Define optimal values for TF_{exp} 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 TF_{exp} in the mouse promoters
Step 5: Define the optimal values for TF_{exp} 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 TF_{exp} target genes

Supplementary Figure 5: **Flowchart of the ChIPModules approach.**

Shown is a schematic indicating the steps needed to develop a database of target promoters for a particular site-specific human transcription factor. 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 (see **Figure 2A**).

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 (see **Figure 2B**).

Step 5: Define the optimal values for the match to the core consensus and PWM for the TF_{exp} in the mouse promoters (see **Figure 2C**). 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 (see **Formula 4 and 5 in Materials and Methods**). Validate the training dataset by ROC method using a range of separation values (see **Figure 4**). 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 (see **Table 3**).

Step 9: Apply the modules to a genome-wide computational analysis of TF_{exp} target genes (see **Supplementary Table S2**).