**ChIP-seq data**

We downloaded all ChIP-seq data with defined narrow peaks and with restriction date until March 11, 2010 from:

<http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC/wgEncodeYaleChIPseq/>.

For each ChIP-seq dataset, we used the cutoff FDR<0.01 (q>2) to extract peaks from the narrow-peak data file. Meanwhile, we also overlapped these peaks with the human non-coding regions we used to predict motif modules. The peaks that do not have any overlap with our region were eliminated. The number of the remaining peaks is in column D of supplemental file 2. The 14 ChIP-seq datasets are in column B of the supplemental file 2.

These ChIP-seq data correspond to 30 motifs. The 30 motifs are in the following: M00037, M00059,M00069,M00118,M00123,M00127,M00128,M00172,M00203,M00223,M00322, M00346,M00492,M00496,M00517,M00615,M00671,M00776,M00777,M00789,M0079, M00799, M00803, M00919, M00920, M00924,M00926,M00983,M01034, M01035.

**Overlapped binding sites**

For every motif, we matched its binding sites to their corresponding ChIP-seq peaks. A binding site is considered to overlap with a peak, if it is within the peak. Then we obtained the overlapped binding sites for every motif. The number of the predicted binding sites for a motif is defined as L. The number of overlapped binding site is defined as N, while the number of peaks is defined as M.

**Threshold of every motif**

For every motif, we collected its predicted binding sites from all significant motif modules. Then we got the sequence of the binding site. We defined the score of a binding site as:

where d is the length of this motif,  is the count of nucleotide b at position i in this binding site,  is the background distribution of the nucleotide b, and Here == 0.2960936, == 0.2039064, which is the nucleotide frequency in human non-coding sequences. For every motif, the minimal score of the predicted binding sites within some ChIP-seq peaks was defined as the threshold (namely T) of this motif. If a segment has a score larger than this threshold, this segment will be called as a putative binding site of this motif under consideration.

**Compare predicted binding sites of individual motifs with individual ChIP-seq experiments**

For every motif, we extracted the sequences of all peaks, and calculated the score of every segment and its reverse complement segment in these sequences similarly as what we did for the predicted binding sites. The number of peaks that have at least one binding site with the score larger than T was defined as K. These binding sites were thought as true binding sites of this motif, since they had a high score and they show binding activities in the ChIP-seq experiment.

From the above description, the number of real binding sites is K, and the number of predicted binding sites is L. The number of overlapped binding site N is the number of true predictions. So precision=N/L, recall= N/K.

There are 12 columns in the supplemental file 2 for the comparison of binding sites of individual motifs with ChIP-seq experiments.

A: motif name

B: the name of the original ChIP-seq data

C: the corresponding factor

D: number of peaks within the human regions we used.

E: the overlapped binding sites of the motif and the ChIP-seq data (N)

F: the number of binding sites of the motif (L)

G: the threshold (the minimal score of all binding sites of the motif) (T)

H: the number of real binding site (=peaks that have scores larger than the threshold)(K)

I: precision=N/L

J: recall=N/K

K: precision/recall

**Compare CRMs of motif pairs with a pair of ChIP-seq experiments**

There are 32 predicted motif pairs in our predictions that are composed of the 30 motifs mentioned above. Among them, 8 motif pairs correspond to the same factor. The remaining 24 motif pairs are composed motifs from different transcription factors. We thus compared the predicted CRMs of these 24 motif pairs with the corresponding pairs of ChIP-seq data.

We have defined real binding sites above (column H of the supplemental file 2). For a motif pair, we define the real CRMs as the overlapped peaks for the two transcription factors. That is, two peaks defined in column H of the supplemental file for the two transcription factors must overlap in order to claim to be a real CRM. Assume the number of the real CRMs is m. Compare these real CRMs with our predicted CRMs for the same motif pairs, the overlapped CRMs are the true predictions. Assume the number of overlapped CRMs is n. Then n/m defines the recall of the method at the motif pair level. The result is showed in supplemental file 3.

There are 16 columns in supplemental file 3. The first 6 columns are the motif, corresponding factor, corresponding ChIP-seq data, number of predicted binding sites, number of peaks in the ChIP-seq data and the number of overlapped binding sites, for the first motif in the motif pair. The next 6 columns are for the second motif in the motif pair, in the same format as the first 6 columns. The 13-th column is the number of predicted CRMs containing the motif pairs, the 14-the column is the number of "true CRMs"(m, these binding sites are the overlap of the binding sites in the 6-th and the 12-th column), the 15-the column is the number of true CRMs overlapped with the predicted CRM(n,these binding sites are the overlap of the binding sites in the 13-th and the 14-th column), and the last culumn is n/m.