# Supplementary Information

## DNase I hypersensitive sites (DHS)

We downloaded the previously identified DHSs of 19 human cell lines from UCSC genome browser (http://genome.ucsc.edu/) [1], representing a wide variety of human tissues. The sequencing raw data were aligned to the human reference genome (built hg.19) using BWA and were then smoothed using a kernel density estimator, F-seq [2]. Then, DHS peaks were identified as having a -log10 (*P-value*) >= 1.3. All sites of DHS peaks can be downloaded from our website: http://donglab.ecnu.edu.cn/data/DHS/index.html

## DHSs annotation

DHSs were classified according to the genomic regions of genes. If a DHS located in the transcription start site region (TSS) of any transcript isoforms of a gene, it was classified as a TSS DHS for the focal gene. Those DHSs were classified as gene body DHSs if they overlap with any regions of the exons or introns, and all other DHSs which do not overlap with any region of a gene were classified as intergenic DHSs. Then we establish the associations between DHSs and genes. TSS DHSs and gene body DHSs were associated with the genes they overlapped. For each intergenic DHS, we used BEDTOOLS software [3] to find the nearest gene, and associated it with that DHS if the distance between them was less than 200kb.

## miRNA targets prediction

The miRNA targets were taken from three previously published *in silico* miRNA target prediction methods, including TargetScan (http://www.targetscan.org version 5.1) [4], PITA (http://genie.weizmann.ac.il/pubs/mir07/mir07data.html) [5] and Pictar (http://genome.ucsc.edu four-way) [6]. Those miRNA targets predicted by TargetScan with a total context score < -0.3 were removed, and those with at least one conserved 7-mer or 8-mer were chosen as reliable miRNA targets. For PITA targets, a score less than -10 was selected as the threshold to choose reliable miRNA targets. To minimize

the false positive of miRNA target prediction, a high-quality miRNA target data set was generated by intersecting data generated by at least two different *in silico* miRNA target prediction methods. Those without being detected by any methods were defined as miRNA non-targets. Furthermore, we also downloaded experimentally verified miRNA target data from miRTarBase database (http://mirtarbase.mbc.nctu.edu.tw). The data can be downloaded from our website http://donglab.ecnu.edu.cn/data/DHS/

**Transcription factor binding site**

We totally downloaded 229 ChIP-seq datasets from UCSC genome browser, representing the DNA footprints of 59, 120, 86 and 64 TFs in four human cell lines (H1, K562, GM12878 and HepG2), respectively. A total of 282982, 689191, 593813 and 589960 ChIP-seq peaks were found in H1, K562, GM12878 and HepG2 cell lines, respectively. The ChIP-seq peaks was defined as TFBS in our work. All data can be downloaded from our website http://donglab.ecnu.edu.cn/data/DHS/

**PWM Scan**

We downloaded 789 PWMs from the JASPAR, TRANSFAC and Uniprobe databases, which represents vertebrate TFs. "PWM Score'' is a log-likelihood ratio of the probability of a given sequence under the PWM model, compared to a random sequence model. Each TF is represented by a specific PWM (a matrix of frequencies) with which this TF is expected to bind certain DNA motif. Each PWM was used to score the intergenic, TSS and gene body DHS regions while looking for subsequences that closely match the binding motif represented by the PWM. Next, we scanned the sequence from each DHS and non-DHS regions. For each location, a score was calculated based on the probability that the sequence was generated in PWM model versus the probability that the specific sequence was generated in background model. The first-order Markov Model trained on a 500-bp window centered at the base pair was applied in the background model. This method could effectively correct for the underlying dinucleotide composition and separate signal from noise. The scores were generated for each base pair. A 60-bp sliding window was moved across the sequence,

and we summed scores of all base pairs in each window. The maximum window score was determined as the TFBS score for that TF. This sliding window based method could account for local clustering of binding sites, which have been shown to be more likely to be bound by TFs than single binding sites. In general, one gene may be associated with more than one DHSs, and these regions were assumed to be the putative regulatory region for that gene. Here, we assigned the maximum TFBS score of that region to each gene.

**Support vector machine (SVM)**

LibSVM package [7] was employed to construct the SVM model and evaluated the performance of the models using five-fold cross-validation. The process of SVM could be summarized as the following steps:

1. Dataset preparation

   A data set matrix should be construct under the following format,

$$\begin{bmatrix} y_1 & x_{1,1} & x_{1,2} & x_{1,3} & \cdots & x_{1,m} \\ y_2 & x_{2,1} & x_{2,2} & x_{2,3} & \cdots & x_{2,m} \\ y_3 & x_{3,1} & x_{3,2} & x_{3,3} & \cdots & x_{3,m} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ y_{n-1} & x_{n-1,1} & x_{n-1,2} & x_{n-1,3} & \cdots & x_{n-1,m} \\ y_n & x_{n,1} & x_{n,2} & x_{n,3} & \cdots & x_{n,m} \end{bmatrix}$$

   Where n represents the number of objects, m represents the number of variables, y represents the binary vector (represent two categories) and x represents the variables. For object *i*, $y_i$ represents the class of object *i* and $x_{i,j}$ represents the value of variable *j* in the object *i*.

2. Model construction based on training set

   The SVM model construction can be simplified as an optimization problem:

$$\text{Minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$

$$Subject\ to \quad y_i\left(w^{'}x_i + b\right) \geq 1 - \xi_i$$

   where w is the coefficients of the hyperplane, C is the penalty coefficient and $\xi_i$ is

the slack variable.

The dual of this optimization problem is:

$$\text{maximize} \quad W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x, x_j)$$

$$subject\ to \quad C \geq \alpha_i \geq 0, \sum_{i=1}^{n} \alpha_i y_i = 0$$

Where w is recovered as $w = \sum_{j=1}^{S} \alpha_{t_j} y_{t_j} x_{t_j}$, S is the support vector sets. *K*(*,*) is the kernel function, i.e. polynomial kernel, radial basis function kernel, sigmoid kernel, etc.

3. Model evaluation using testing set.

For a test object z, the discriminate function essentially is a weighted sum of the similarity between z and a preselected set of objects (the support vectors),

$$f(z) = \sum_{x_i \epsilon S} \alpha_i y_i K(z, x_i) + b$$

,

where S represents the support vector sets.

Based on the miRNA-mRNA relationships, genes can be classified into miRNA targets and non-targets. TFBS scores were used for SVM classifiers to discriminate miRNA targets and non-targets. Data matrix could be shown under the following format

$$\begin{bmatrix} y_1 & x_{1,1} & x_{1,2} & x_{1,3} & \cdots & x_{1,m} \\ y_2 & x_{2,1} & x_{2,2} & x_{2,3} & \cdots & x_{2,m} \\ y_3 & x_{3,1} & x_{3,2} & x_{3,3} & \cdots & x_{3,m} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ y_{n-1} & x_{n-1,1} & x_{n-1,2} & x_{n-1,3} & \cdots & x_{n-1,m} \\ y_n & x_{n,1} & x_{n,2} & x_{n,3} & \cdots & x_{n,m} \end{bmatrix}$$

Where n is the gene number and m is the TF number, $y_i$ represent the miRNA target status of gene *i*, and $x_{i,j}$ represents the TFBS score of TF *j* bound to gene *i*.

## References:

1. K. R. Rosenbloom, J. Armstrong, G. P. Barber, J. Casper, H. Clawson, M. Diekhans, T. R. Dreszer, P. A. Fujita, L. Guruvadoo, M. Haeussler, R. A. Harte, S. Heitner, G. Hickey, A. S. Hinrichs, R. Hubley, D. Karolchik, K. Learned, B. T. Lee, C. H. Li, K. H. Miga, N. Nguyen, B. Paten, B. J. Raney, A. F. Smit, M. L. Speir, A. S. Zweig, D. Haussler, R. M. Kuhn and W. J. Kent, *Nucleic acids research*, 2015, **43**, D670-681.
2. A. P. Boyle, J. Guinney, G. E. Crawford and T. S. Furey, *Bioinformatics*, 2008, **24**, 2537-2538.
3. A. R. Quinlan and I. M. Hall, *Bioinformatics*, 2010, **26**, 841-842.
4. B. P. Lewis, I. H. Shih, M. W. Jones-Rhoades, D. P. Bartel and C. B. Burge, *Cell*, 2003, **115**, 787-798.
5. M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul and E. Segal, *Nature genetics*, 2007, **39**, 1278-1284.
6. A. Krek, D. Grun, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. MacMenamin, I. da Piedade, K. C. Gunsalus, M. Stoffel and N. Rajewsky, *Nature genetics*, 2005, **37**, 495-500.
7. C. C. Chang and C. J. Lin, 2001.