(2)

- SELEX-seq

SELEX analysis

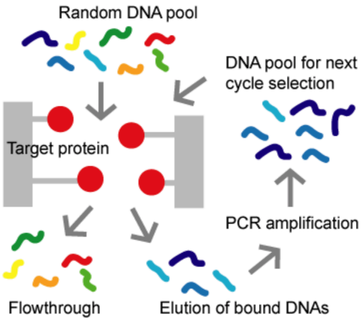


Fig.1 SELEX work flow

The SELEX (**S**ystematic **E**volution of **L**igands by **EX**ponential enrichment) procedure explores the target protein binding DNA sequence through repeated cycles of binding selection and PCR amplification. When Random DNA sample goes through the column has target protein in it, DNAs that has the recognition sequence bind to target protein, and those don't have just flow through. Next PCR the eluted bound DNAs and select again. Repeat this procedure, a DNA pool containing significant information about the recognition sequence will be provided.

In the original SELEX procedure, the random DNA sample could be generated by PCR synthesized oligonucleotides containing random sequence. So it is a *in vitro* Method.

However, genomic SELEX-seq, the sequence library is derived directly from the chromosomal DNA of the target organism. So in genomic SELEX-seq we could also determine the binding position of the target protein on its genome sequence.

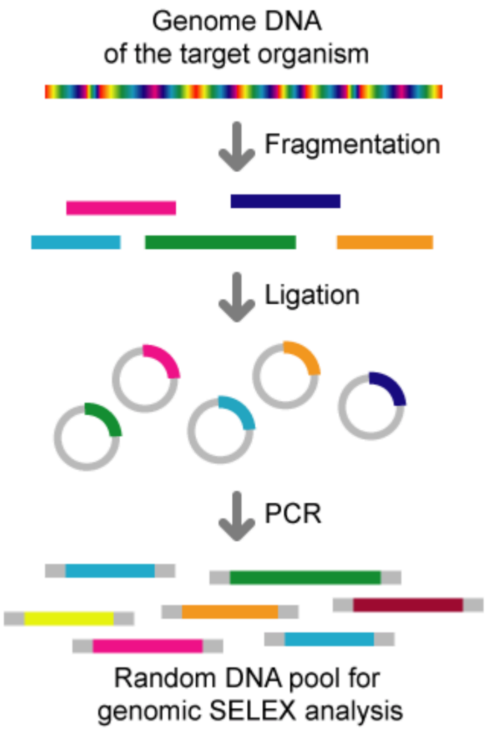


Fig.2 genomic SELEX-seq work flow

After constructing the DNA library, feed it into the SELEX procedure and select the DNA that has the recognition sequence and then do the analysis.

- PBM

Protein Binding Microarrays (PBMs), is a high-throughput method characterizing the *in vitro* DNA binding site sequence specificities of transcription factors.

First, a protein of interest is expressed and purified. Then assaying the protein on a double-stranded DNA microarray spotted with a large number of potential DNA binding sites. The protein-bound microarrays are then washed to remove nonspecifically bound proteins. After that the microarray is labeled by a fluorescent-antibody specific to the target protein and the fluorescence signals give us the information(data) about the DNA binding sites.

The resulting DNA binding site data can be used in a number of ways, including for the prediction of the genes regulated by a given transcription factor, annotation of transcription factor function, and functional annotation of the predicted target genes.

- ChIP-seq

ChIP-seq, is a method used to analyze protein interactions with DNA. ChIP-seq combines chromatin ImmunopreciPitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins.

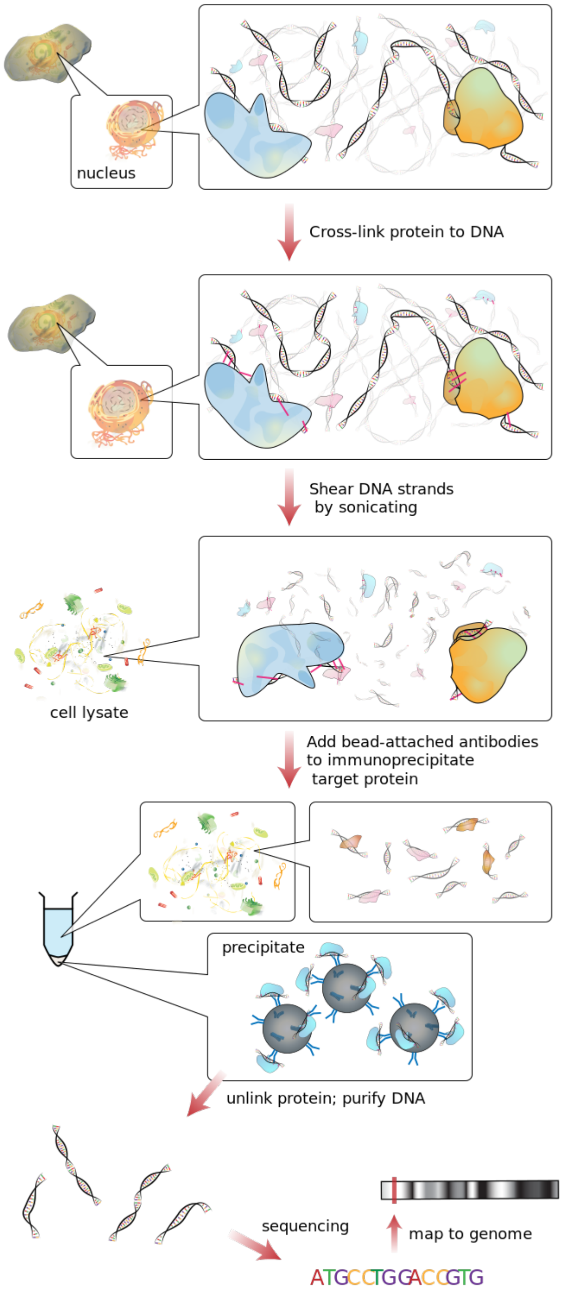


Fig.3 ChIP-seq workflow

First, add chemical agent to cross-link the DNA and protein as it is in the nucleus, so it is an *in vivo* approach. Then fragment the DNA into small pieces by sonication. Adding antibodies to immunoprecipitate target protein. Construct the DNA library for the selected DNAs and go through a high-throughput sequencing procedure. And then we could map the attained reads data to genome and conduct downstream analysis to gain knowledge about the binding sites of the target protein in genome.

(3&4)

Coefficient of determination shown below. Code is in script “Q3.R”.

Table.1 averaged R^2(coefficient of determination) for model “1-mer” and “1-mer+1-shape”

Dataset	Mean(R^2)	Model
Mad	0.7171604	1-mer
Mad	0.7344308	1-mer+1-shape
Max	0.7028587	1-mer
Max	0.7198759	1-mer+1-shape
Myc	0.7264695	1-mer
Myc	0.7304204	1-mer+1-shape

(5) Code is in script “Q5.R”.

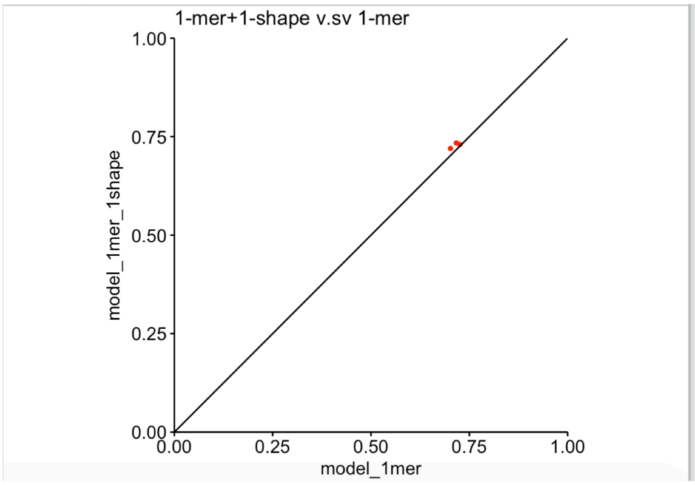


Fig. 3 Comparision of R^2 of model 1-mer+1-shape v.s. 1-mer

(7)

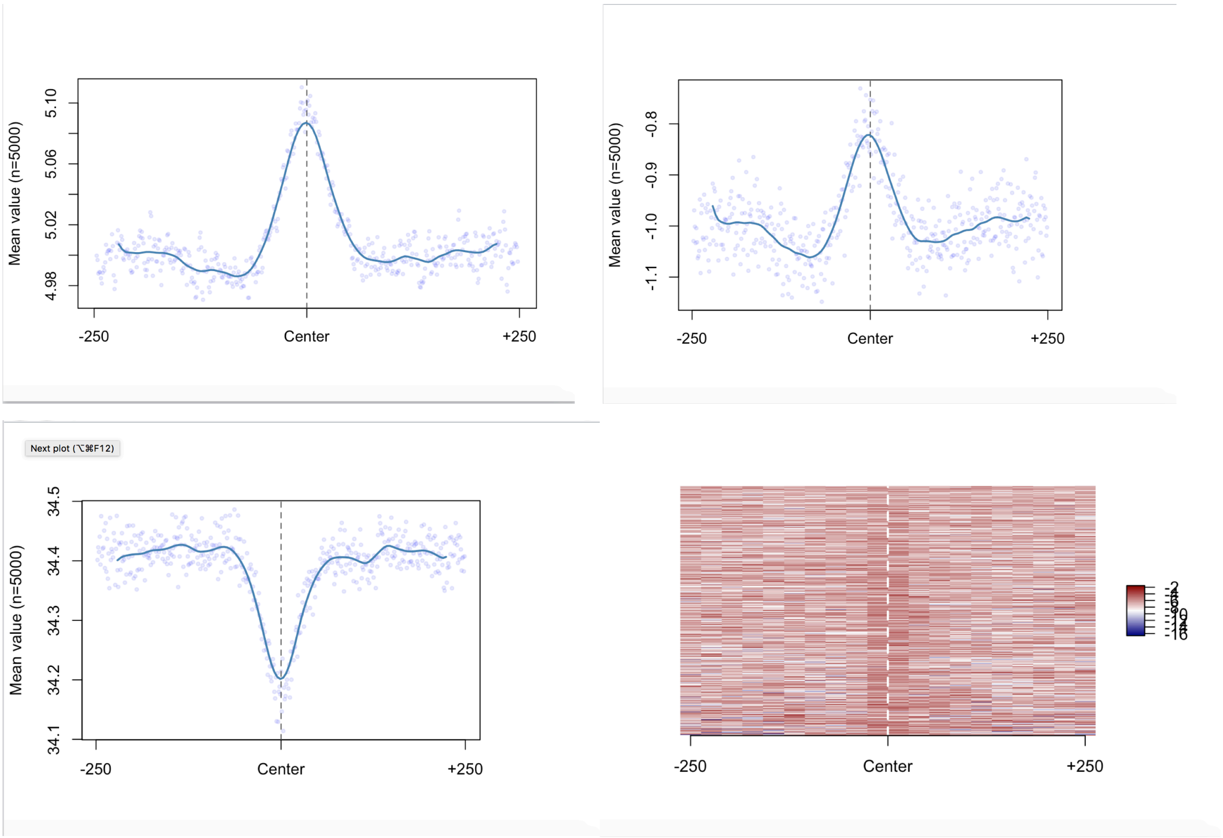
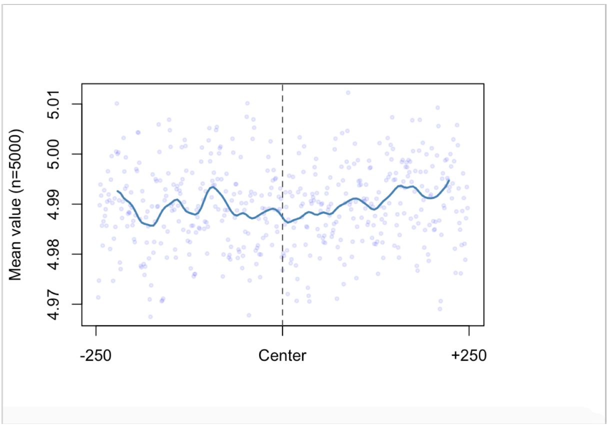
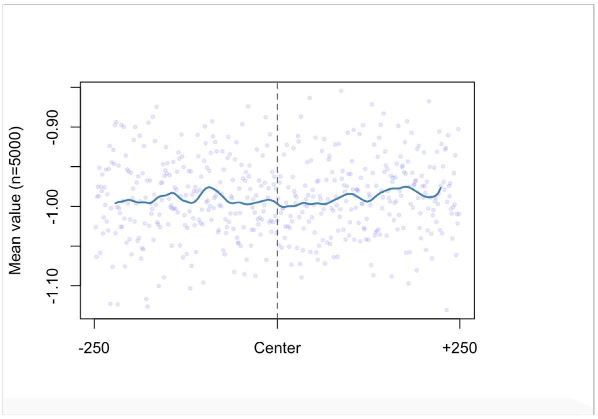


Fig. 4 Shape parameters of bound sequence, from left to right, up to bottom: Minor groowidth(MGW), propeller twist (ProT), Roll, and helix twist(HelT), sequence lengt: 500.





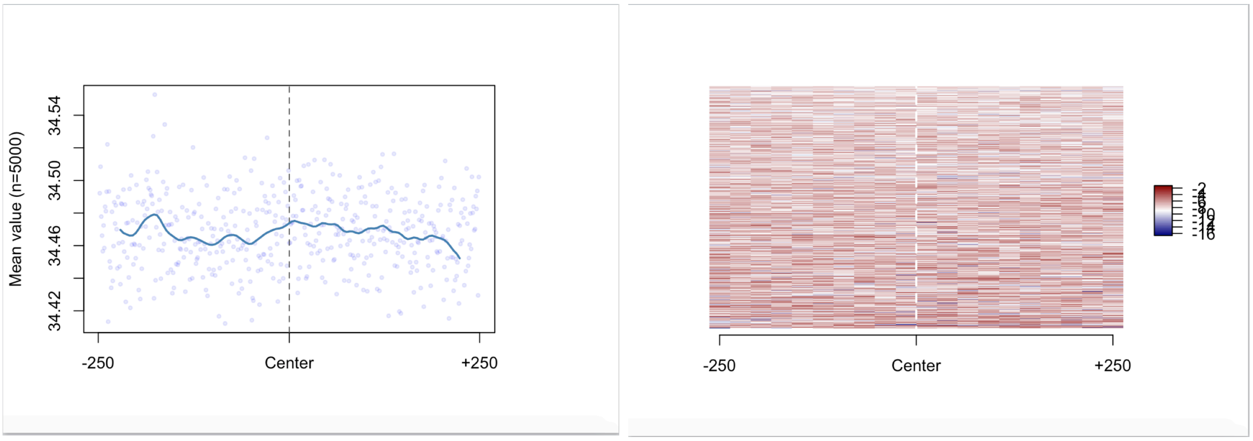


Fig. 5 Shape parameters of unbound sequence, from left to right, up to bottom: Minor groowidth(MGW), propeller twist (ProT), Roll, and helix twist(HelT), sequence lengt: 500.

By comparison, we could tell that the CTCF binding will change the DNA shape and that changes could be identified by ChIP-seq technique: near the center of the binding site, the minor groove width increases by 0.08 to the average, the propeller twist increasesby 0.2 compared to the average, and helix twist decreases by 0.2 to the average.

(8)

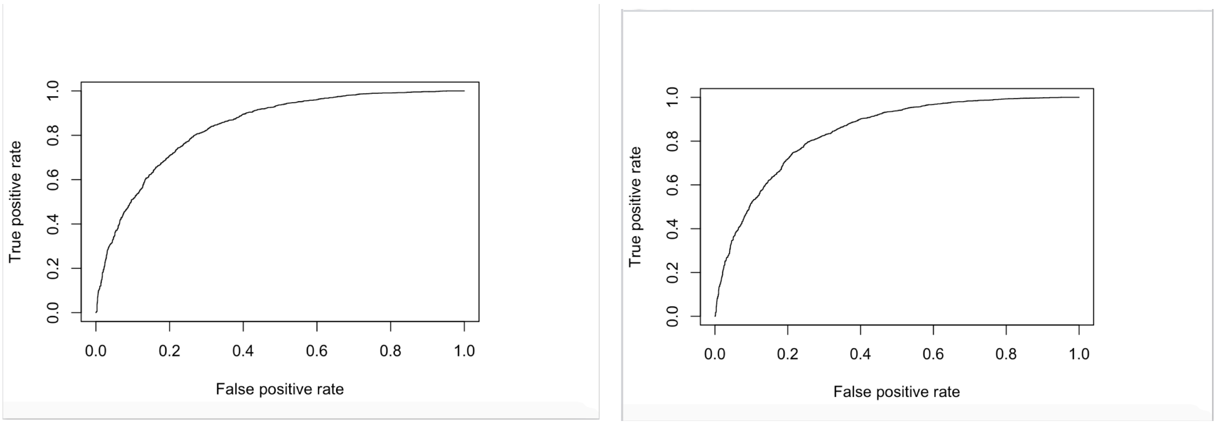


Fig.6 ROC curve of logistic regression of model “1-mer”(right) and modle”1-mer+1-shape”

The AUC(area under the curve) is 0.8419(“1-mer”) and 0.8396(“1-mer + 1-shape”).

This result shows choosing feature vector to be “1-mer” alone or “1-mer+1-shape” has the same prediction performance with respect to CTCF binding. Which may suggest that the DNA sequence play a more important role in the CTCF binding.