ENGINE: an enhancer gene interaction detection algorithm using robust feature extraction

Lou Shaoke

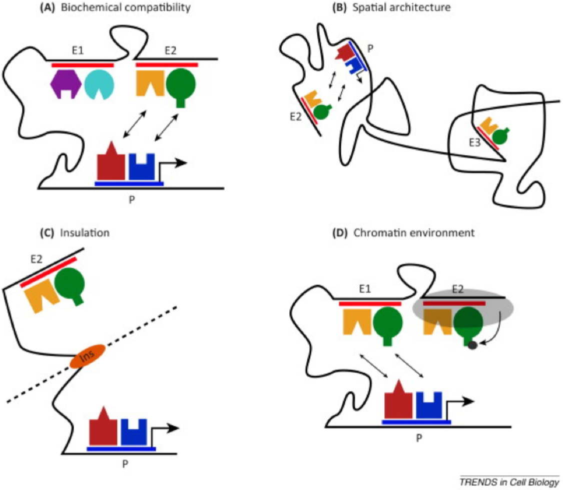
Department of Molecular Biophysics and Biochemistry

loushaoke@gmail.com

February 5, 2016

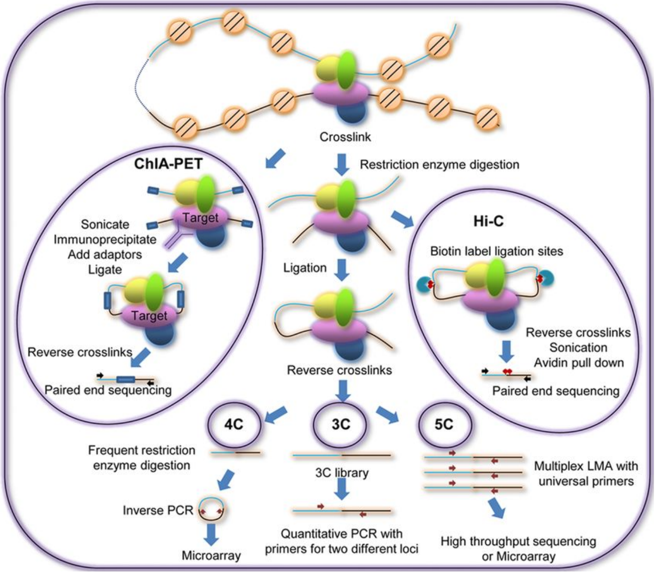
Yale

Biological of enhancer gene linkage

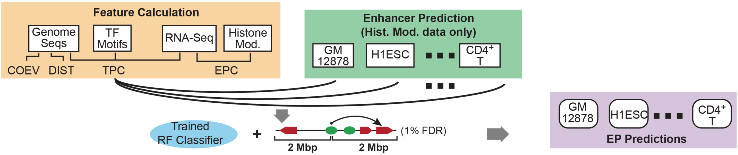


Classic problem: enhancer-promoter interaction. Biological compatiblity(sequence feature and motif); spatial compatibility (3d interaction); local environment (epigenomic marks)

3D genome techniques

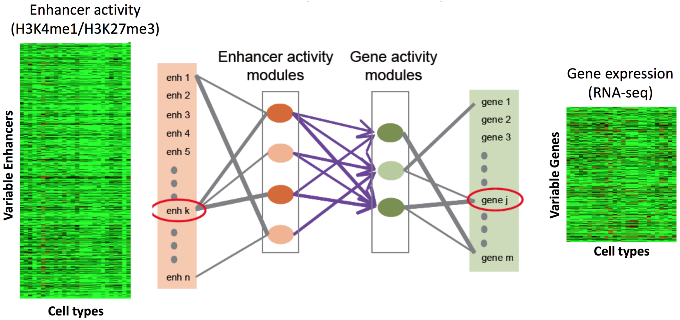


State of the art



IM-PET: Consider information from 3D gnome interactions, DIST(distance) constrain is a triky feature, boosting AUC from 0.7+ to 0.9+.

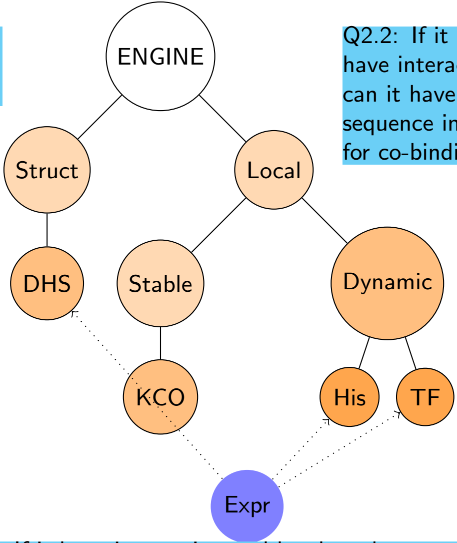
State of the art



LDA:a mixed membership method, didn't use information from 3d genome interaction, and reply on predifined enhancer region, sometimes it has worse aggreements.

Model

Q1: Can DHS indicate putative interaction region



Q2.2: If it possible to have interaction, can it have specific sequence information for co-binding

Q2.1: If it possible to have interaction, can it have epigenetic signal for transcriptional activity

Q3: If it have interaction and local markers, can it lift-up(or be high correlated with) certain transcripts target

Dataset

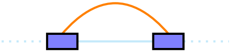
ChIA-PET dataset: K562, Mcf7, Gm12878 and Hela

Gene expression data: Encode TSS based

Open chromatin data: K562, Mcf7, Gm12878 and Hela

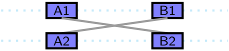
Histone modification and TF data: Ep300, CpG, H3K27ac, K3k4me1, H3k4me3

Positive dataset:

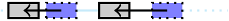


ChIA-PET interaction pairs overlap with mix-membership defined enhancer-gene linkage(from MIT);

Negative dataset:



1) Shuffled one side of interactor regions;



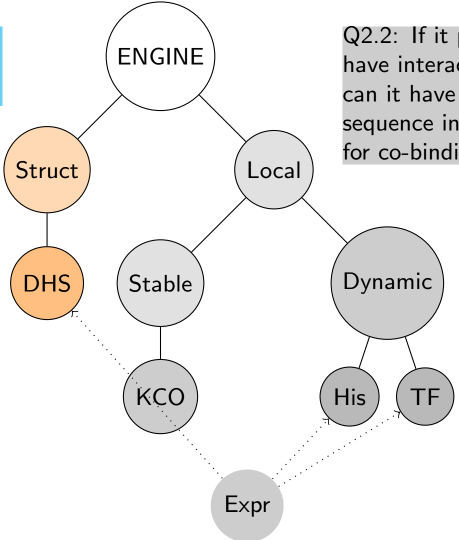
2.1) random shift interaction region

2.2) cell-specific interaction region

DHS and structural information

Q1: Can DHS indicate putative interaction region

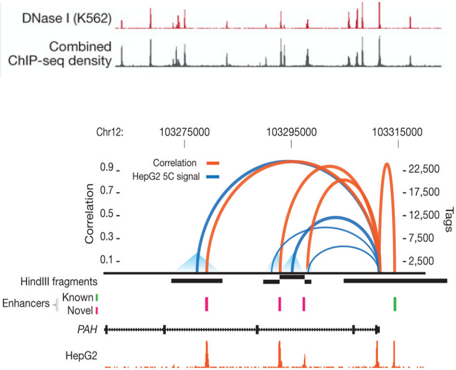
Q2.2: If it possible to have interaction,
can it have specific sequence information for co-binding



Q2.1: If it possible to have interaction,
can it have epigenetic signal for transcriptional activity

Q3: If it have interaction and local markers,
can it lift-up(or be high correlated with) certain transcripts target

DHS: Open chromatin signal

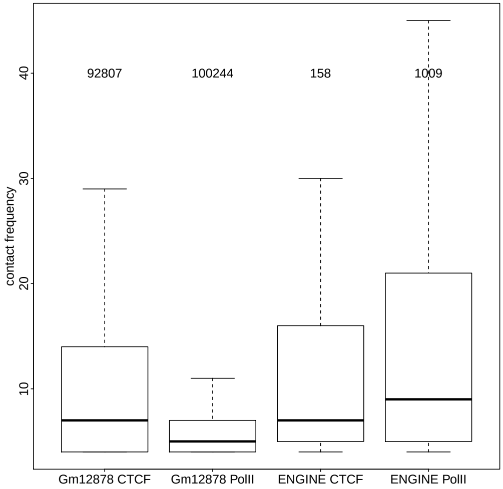


(Robert ET, et al. Nature 2012)

DHS shows high correlation with combined ChIP-Seq signal(0.7+);

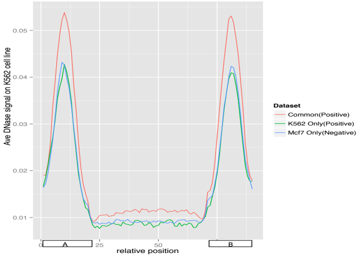
DHS include distal regulatory regions.

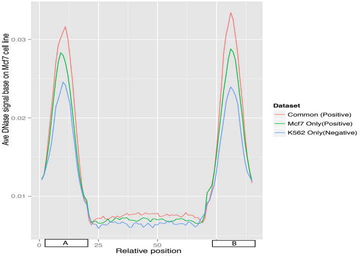
Contact frequency of positive dataset



Positive set are more enriched in PolII-based interaction,and tend to have higher contact frequency.

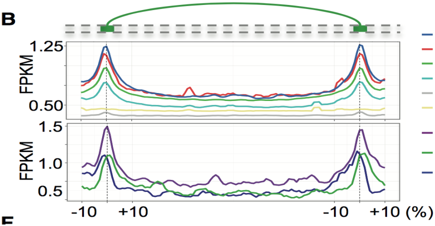
DHS pattern around interaction regions



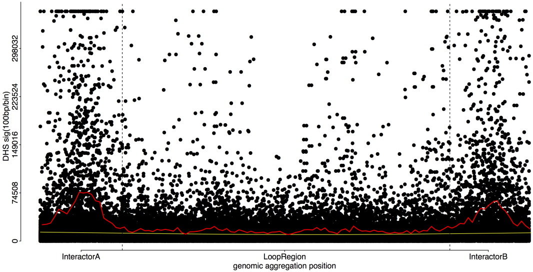


Histone marks and PolII shows similar pattern as the DHS signal

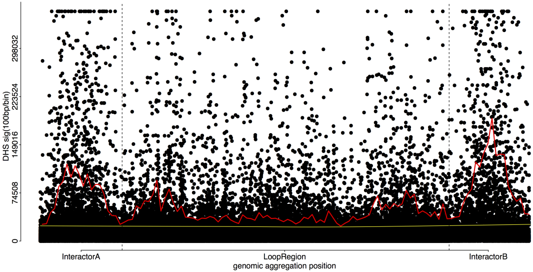
(Tang, et. al. Cell 2015)



Detailed DHS signal distribution

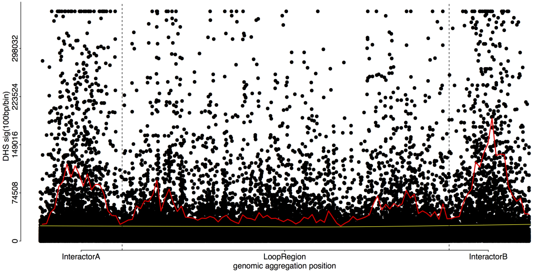


K562 positive interaction

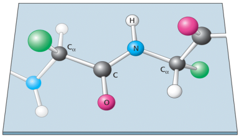


Shuffled K562 interaction

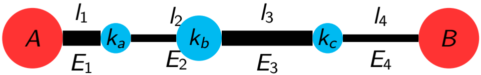
Detailed DHS signal distribution



Shuffled K562 interaction



Molecular flexibility



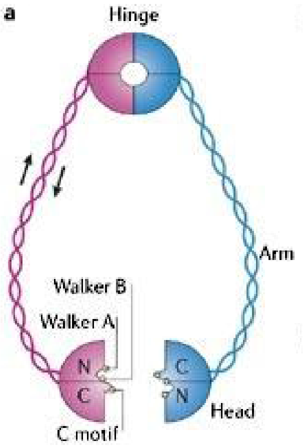
l_i: length of interval region between peaks (bond length)

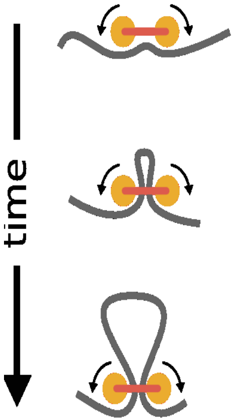
~~E_i: average DHS signal (chromatin flexibility) (bond strength)~~

k_i: high DHS region(peaks), with parameter region size(s_i) and
signal strength (w_i)

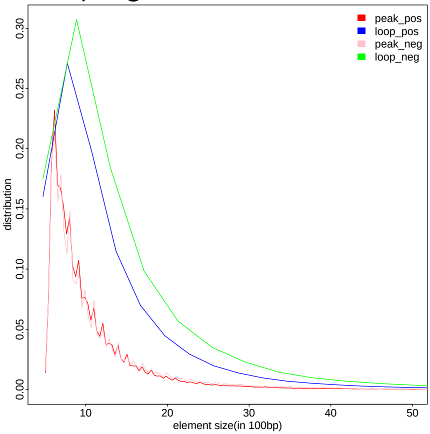
Whether A and B have interaction is determined by l_i, E_i, k_i etc.

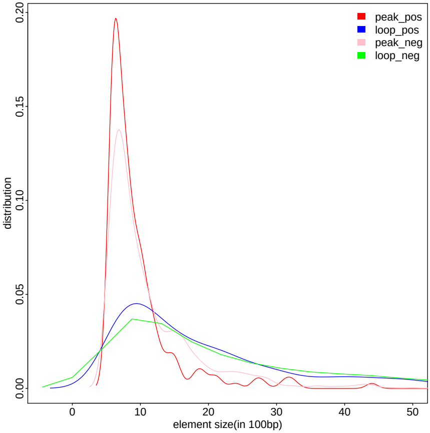
Loop extrusion





Loop regions





Sampled 4000 Gm12878 interactions

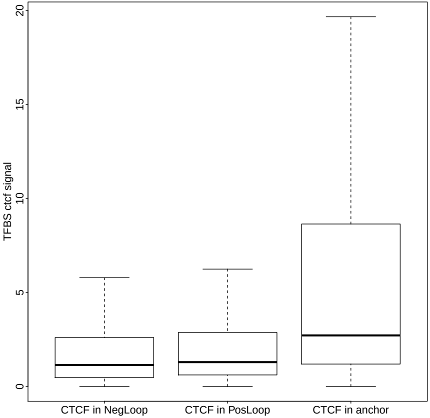
ENGINE interactions

Not all the interaction has transcriptional activity. The loop region should affect the form of 3d interaction.

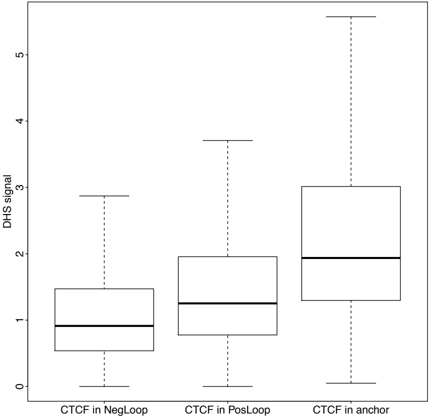
peak: the continous region with signal greater than local mean, loop: the region between two peaks.

Deactivated CTCF binding in the loop region

The shuffled negative set is significant larger than positive dataset



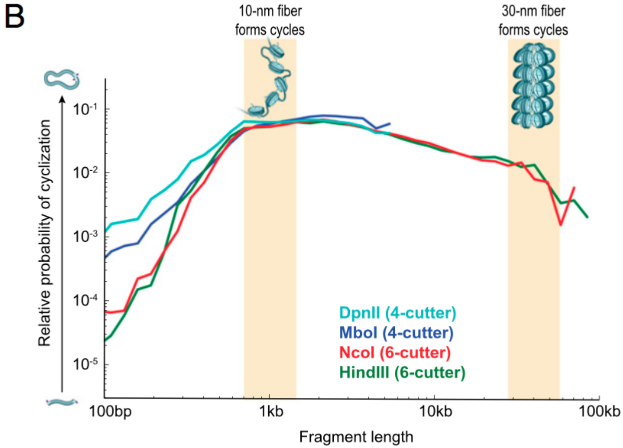
Loop region CTCF are deactivated



CTCF motif in positive loop region has relative lower DHS signal than anchor but higher than negative loop region

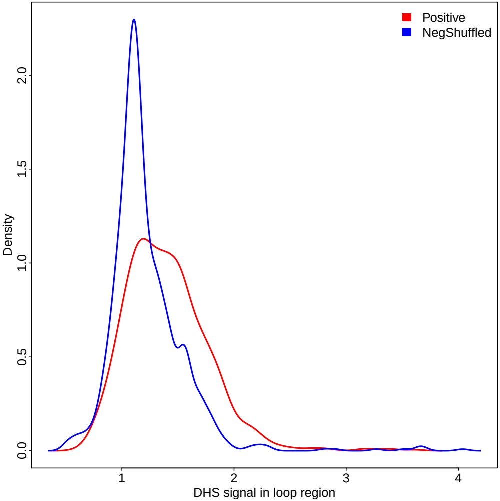
Chromatin Stiffness and DHS signal

(Sanborn et. al. PNAS 2015)



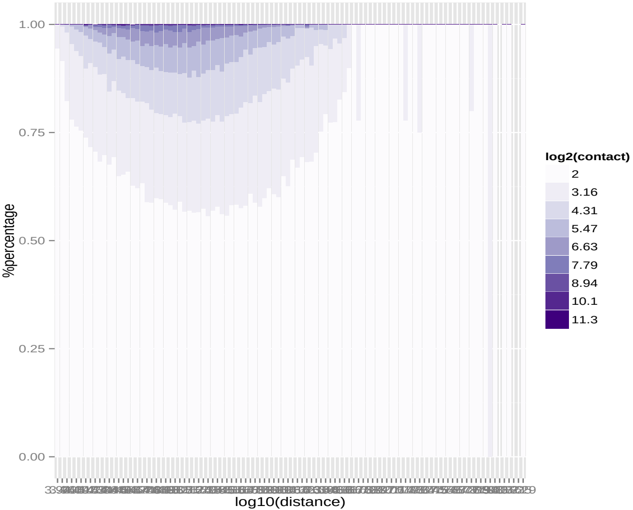
Chromatin too flexible or stiff are both not good for the 3d interaction

Chromatin Stiffness and DHS signal



Relative more stiffness for negative dataset for it covers some high dense chrom regions: Over 30kp with very low/zero DHS signal that might form 30nm fiber. **Can 3d interaction region span so long distance?**

Chromatin Stiffness and DHS signal



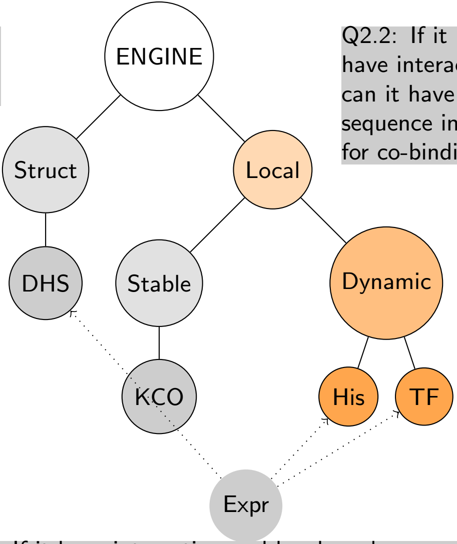
Most of long distant interaction has very low contact frequency. If there are physical interaction, it should have the same chance to get same level frequency. The contact frequency reflect higher level structure by loop extrusion and spacial proximity of the distal interaction, or experimental bias by false joining the blunt end.

Some conclusions for 3d interactions

Suitable stiffness chromatin structure: 10nm fiber(1kb) vs 30nm fiber(30kb), there are no 30nm fiber found in positive set, but 40/909 in negative dataset.
Suitable distance: $D \sim L \times w_c$, L is the real length of loop region(bp), w_c is compact factor, determine compression ratio and chromatin state(10nm or 30nm fiber)
Deactived CTCF motif: most CTCF motif in positive loop region are deactivated.
Very long distance interactions are fake physical interactions?
A: 1) Very low contact frequency, 2) CTCF binding last long time, the loop extrusion is balance of chromatin stiffness and distance, also the energy used for walking along the genome.
Definition of negative dataset are also important.

Dynamic local features

Q1: Can DHS indicate putative interaction region



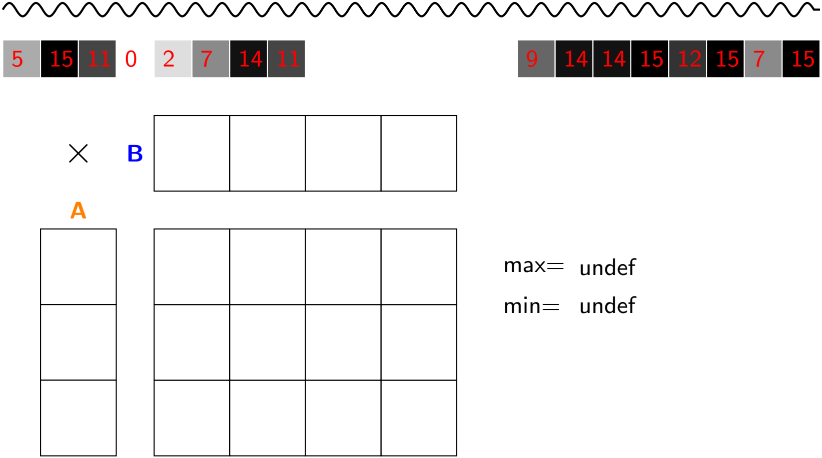
Q2.2: If it possible to have interaction, can it have specific sequence information for co-binding

Q2.1: If it possible to have interaction, can it have epigenetic signal for transcriptional activity

Q3: If it have interaction and local markers, can it lift-up(or be high correlated with) certain transcripts target

Dynamic feature selection

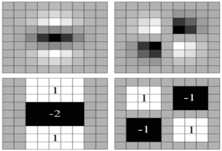
- The tissue-specific enhancer region is hard to predict, chromHMM and CAGE data. The identification and mechanism of enhancer are still not well studied
- The interaction region is not precisely defined, covering a vast non-regulation associated region. However, the average signal over the whole region may bias the results
- The interation involve two different chromatin region, and the close region need some consistent regulatory patterns to exhibit transcriptional activity.



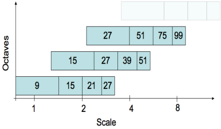
max= undef

min= undef

SURF (Speeded Up Robust Feature) is a robust image blob detector and descriptor, first presented by Herbert Bay et al. in 2006.

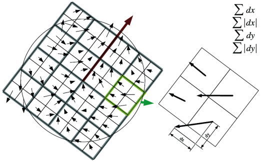


$$H(x,\sigma) = \begin{bmatrix} L_{xx} & L_{xy} \\ L_{xy} & L_{yy} \end{bmatrix}$$



up-scaling filters and scale space

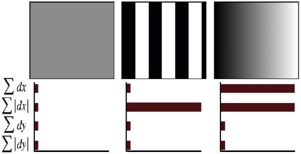
Detection



Descriptor based on Sum of Haar Wavelet Responses

$$v = \{\sum d_x, \sum |d_x|, \sum d_y, \sum |d_y|\}$$

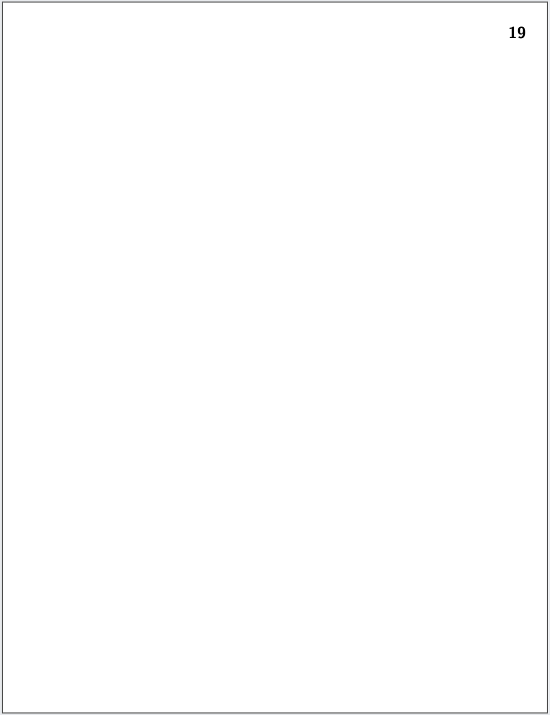
Description



Distance= $\sum (v_1 - v_1')^2$

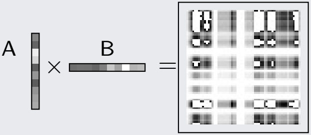
Comparison

Flowchart



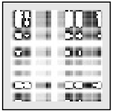
19

Flowchart



408 positive set:K562 ChIA-PET intersect with MIT mix-membership prediction
408 negative set:MCF7 specific ChIA-PET interactions

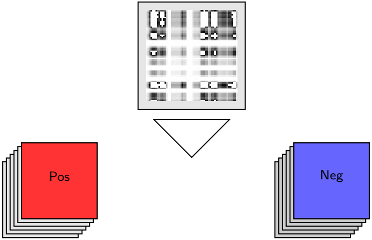
Data transformation



Flowchart

SURF: Speeded Up Robust Features, merits:

- Scale and image rotation invariant detectors and descriptors.
- blob detection
- ...

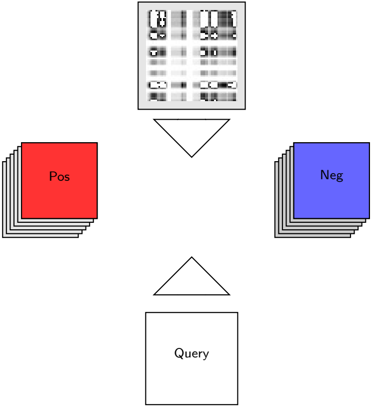


Flowchart

Feature S_i in $N_{j,k}$ matrix (feature sets), and recognition matrix

$$R_{i,j} = \begin{cases} 1, & \text{if } s_i = n_j \\ 0, & \text{otherwise} \end{cases} \qquad (1)$$

The significantly enriched features in positive and negative dataset are defined using hypergeometric test.

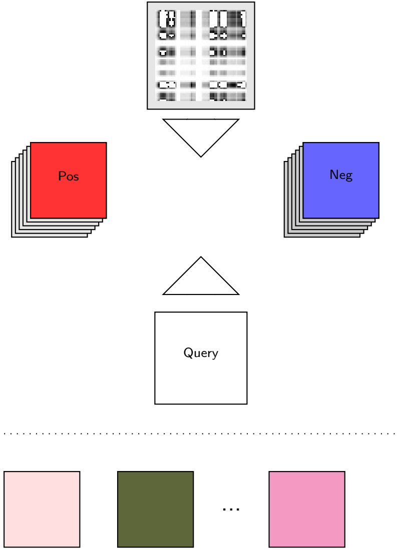


Flowchart

Feature S_i in $N_{j,k}$ matrix (feature sets), and recognition matrix

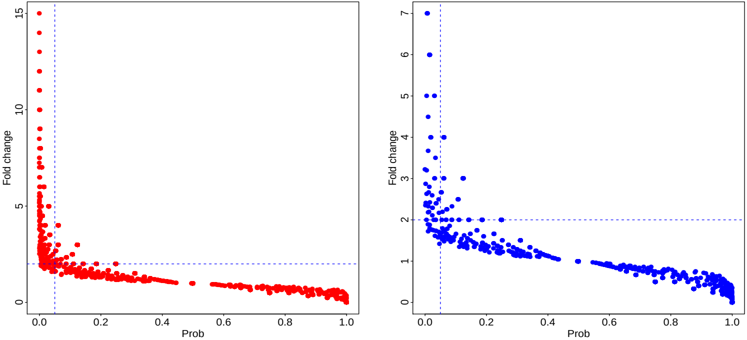
$$R_{i,j} = \begin{cases} 1, & \text{if } s_i = n_j \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

The significantly enriched features in positive and negative dataset are defined using hypergeometric test.



H3k27ac,H3k4me1,H3k4me2,H3k4me3,H3k9ac,
H3k9me1,H3k9me3,P300

Feature selection

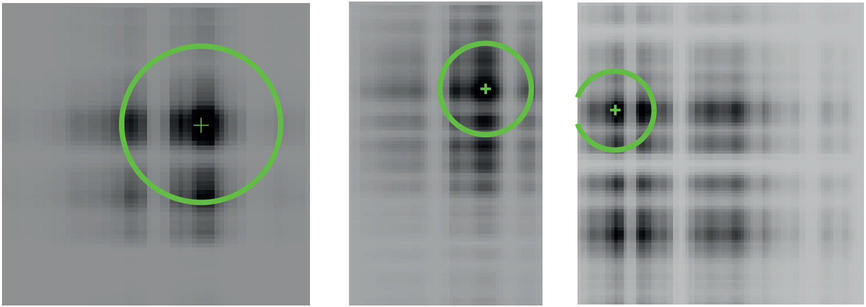


pvalue($= \sum(dhyper(pos\_hit : total\_hit, \#pos\_sample, \#neg\_sample, total\_hit))) < 0.05$ and FC>2, #pos_features in each marker:

H3k27ac	H3k4me1	H3k4me2	H3k4me3	H3k9ac	H3k9me1	H3k9me3	P300	nCpG
395	835	742	462	400	1427	2110	672	1228

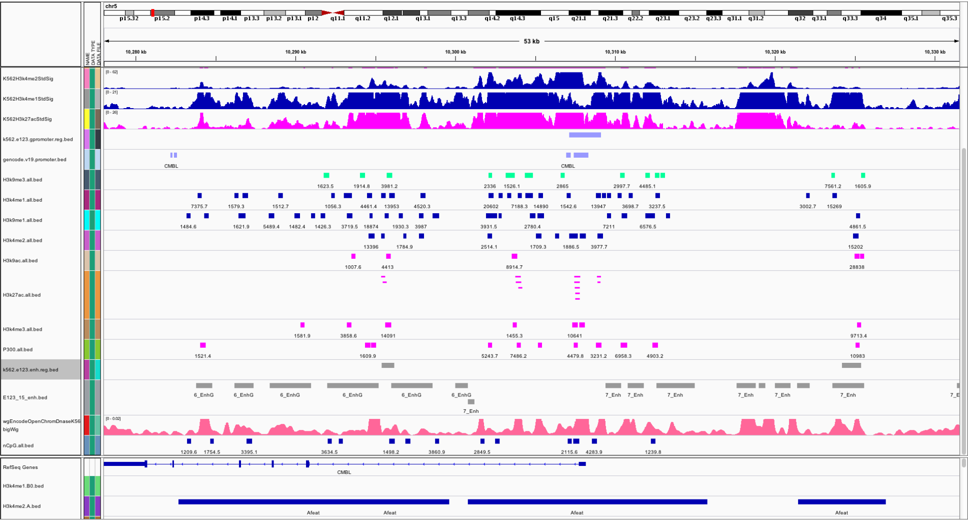
More #sig_features $\neq$ high importance;

Feature visulization

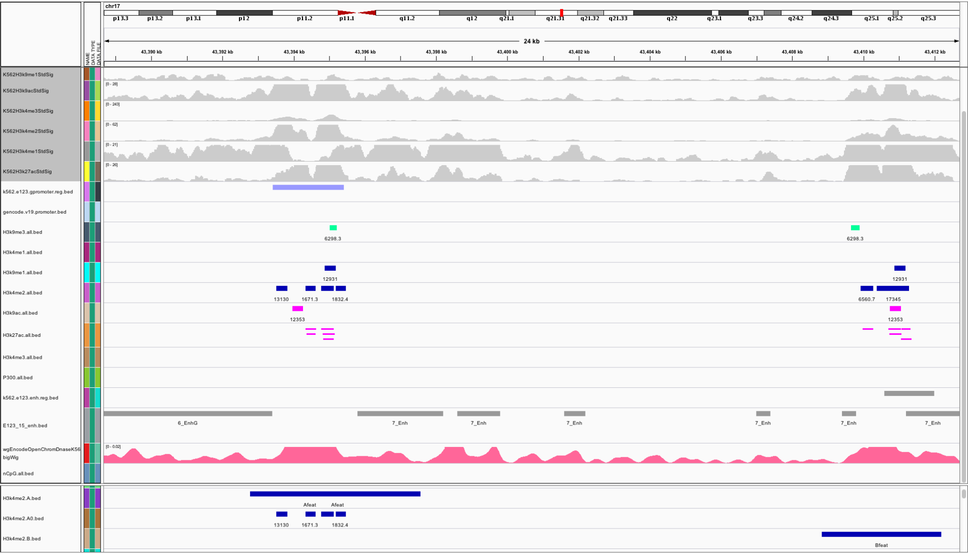


Example for top H3K27ac features

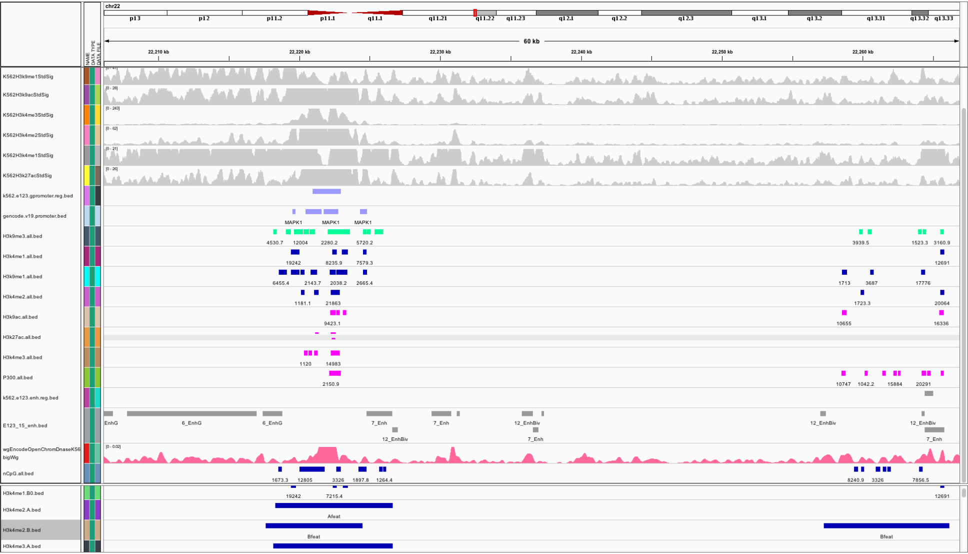
Feature visulization



Feature visulization

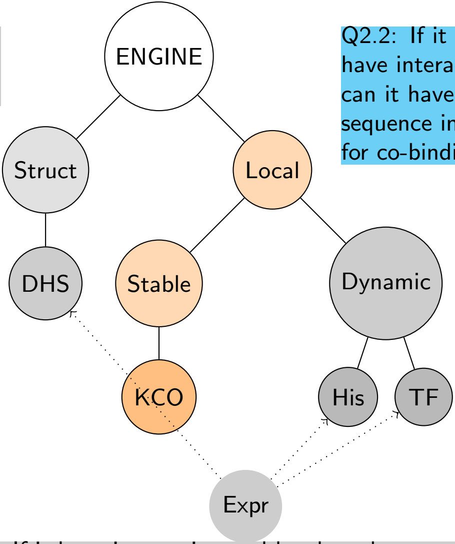


Feature visulization



Local sequence features

Q1: Can DHS indicate putative interaction region

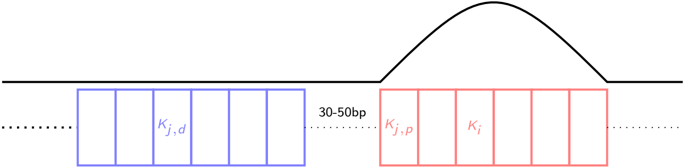


Q2.2: If it possible to have interaction,
can it have specific sequence information for co-binding

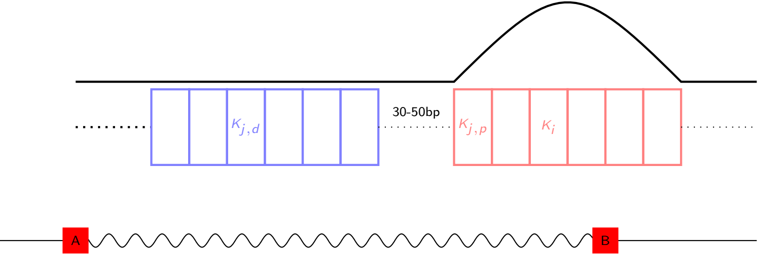
Q2.1: If it possible to have interaction,
can it have epigenetic signal for transcriptional activity

Q3: If it have interaction and local markers,
can it lift-up(or be high correlated with) certain transcripts target

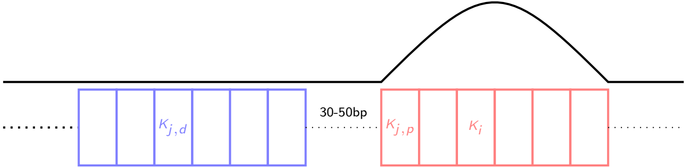
Defined the kmer co-occurence as for any $for \forall i, j \in \{1 \dots n\}$, k_i and k_j is the k-mer ended at the position i or j. $p(k_{j,proximal}|k_i)$ and $p(k_{j,distal}|k_i)$.

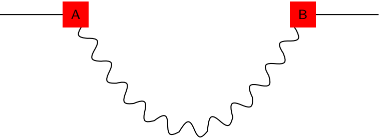


Defined the kmer co-occurence as for any $for \forall i, j \in \{1 \dots n\}$, k_i and k_j is the k-mer ended at the position i or j. $p(k_{j,proximal}|k_i)$ and $p(k_{j,distal}|k_i)$.

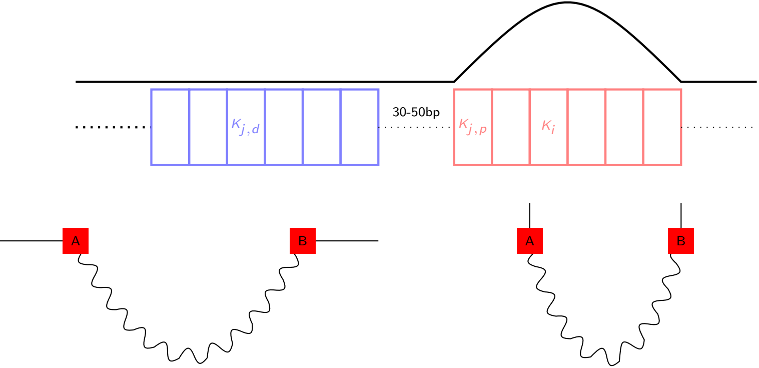


Defined the kmer co-occurence as for any $for \forall i, j \in \{1 \dots n\}$, k_i and k_j is the k-mer ended at the position i or j. $p(k_{j,proximal}|k_i)$ and $p(k_{j,distal}|k_i)$.

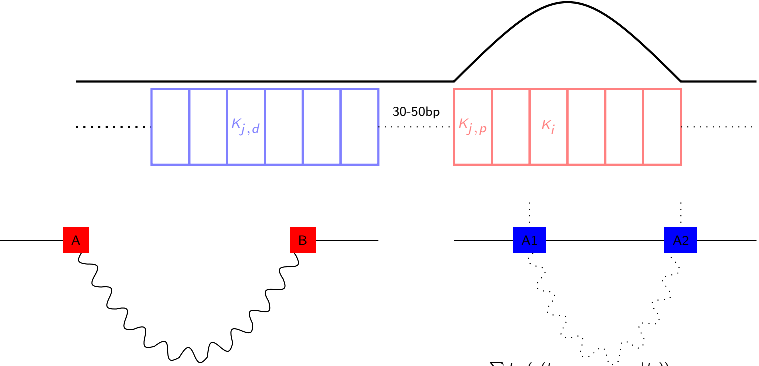




Defined the kmer co-occurence as for any $for \forall i, j \in \{1 \ldots n\}$, k_i and k_j is the k-mer ended at the position i or j. $p(k_{j,proximal}|k_i)$ and $p(k_{j,distal}|k_i)$.



Defined the kmer co-occurence as for any $for\forall i,j \in \{1 \dots n\}$, k_i and k_j is the k-mer ended at the position i or j. $p(k_{j,proximal}|k_i)$ and $p(k_{j,distal}|k_i)$.



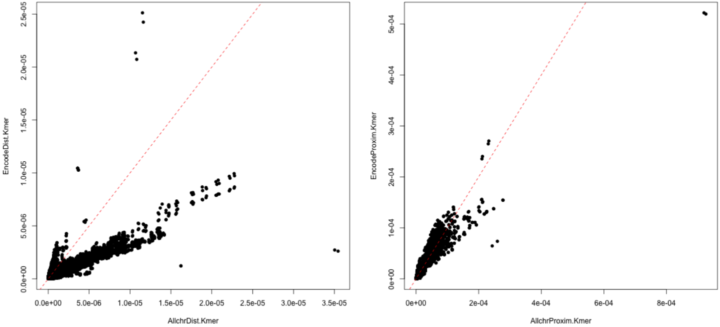
$$KCO_{j,i} = \frac{\sum log(p(k_{j,distal_{encode}}|k_i))}{\sum log(p(k_{j,distal_{background}}|k_i))}$$

We hypothesis, the interaction between paired peaks from interactor A and B, kmers from these peaks might have high chance to present in a distal region(30bp-50bp).

proximal:[-18,18]bp; distal:[-50, -32] bp and [32,50]bp;

Kmer co-occurrence

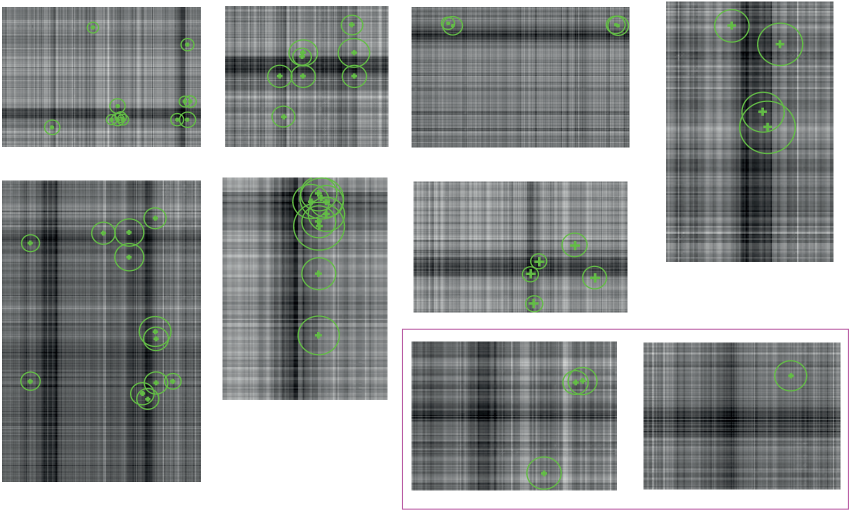
Use All encode enhancer peaks (from FunSeq2) and whole genome sequence(as background), we calculated proximal(-18bp,+18bp] and distal [-50, -32] bp and [32,50]bp co-occurrence frequence $p(k_{j,distal|proximal_{encode}}|k_i)$ vs $p(k_{j,distal|proximal_{background}}|k_i)$.



Proximal region share the co-occurrence pattern but the distal engion quite different between encode peak region and whole genome background.(red dot is y=x).

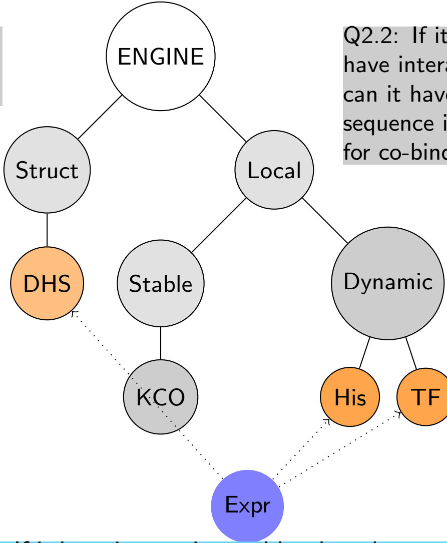
Kmer co-occurence with epigenome signatures

EP300 and H3k27ac significant feature overlap with kmer co-occurrence features



Expression level as the additional information

Q1: Can DHS indicate putative interation region

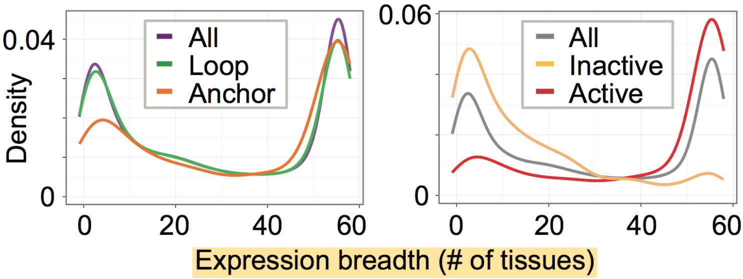


Q2.2: If it possible to have interaction,
can it have specific sequence information for co-binding

Q2.1: If it possible to have interaction,
can it have epigenetic signal for transcriptional activity

Q3: If it have interaction and local markers,
can it lift-up(or be high correlated with) certain transcripts target

Gene expression information



Expression changes for activated gene

Model

Model learning

Importance of features

relationship between structural and local features

relationship between stable and dynamic features

relationship of dynamic features

How variants affect gene linkage

genome wide prediction of gene linkage

Genome wide effect prediction for variants that affect enh-gene linkage

Acknowledgement