Gene linkage

Lou Shaoke

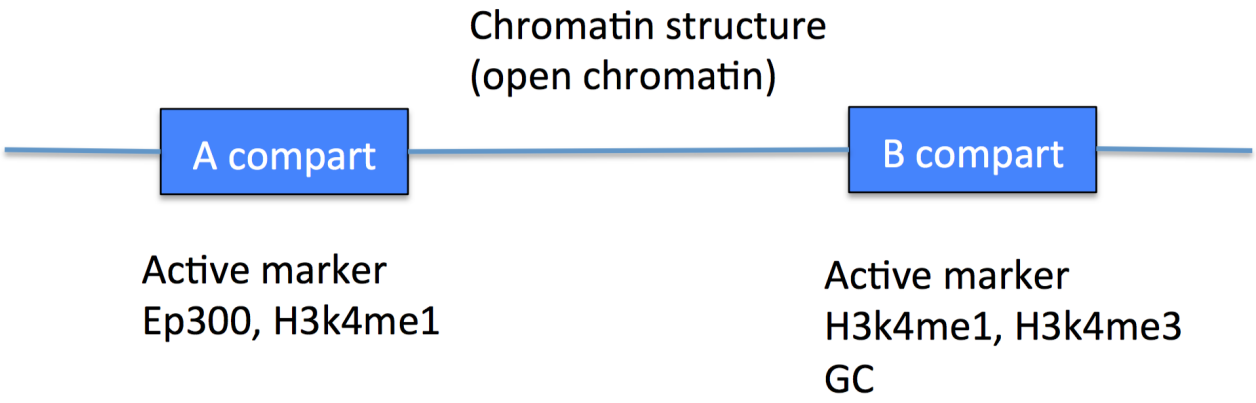
Department of Molecular Biophysics and Biochemistry

loushaoke@gmail.com

September 2, 2015

Yale

Features determine the cobinding



co-binding active region already exists, but regulated by other factors: H3K4me1, EP300, other competitive factors.
We have enhancer-gene pairs, however, the region is quite large(2K or more), however, cobinding region may limited a region less than 100bp or even shorter.
DNase footprinting is also a good signatures to infer TF binding, but not very reliable, and people usually use the hotspot region to intersect FIMO identified binding regions.

Yale

Framework to define co-binding

Why need to do physical interaction?
1) hot region has multiple interaction locus
2) sometimes the region from ChIA-PET still very large
Two steps:

- Step I. Define target binding region pairs for A,B
- Step II. Define partnership for binding regions.

The characteristics of TBR:

- identify short highly conseved region in the current enhancer region(less than 100bp);
- These region has different sequence features compared with its flanking region
- These region are bound by motif/specific complementary mechanisms. The paired co-bound region is not uniq, which means, we can find the similar sequence pattern but stay closely with each other. (This is a strong restrictions)

Yale

Kmer and its derivatives

Composition vectors:
$p(a_1 a_2 a_3 \ldots a_k) = \frac{p(a_1 a_2 \ldots a_{k-1}) \times p(a_2 a_3 \ldots a_k)}{p(a_2 a_3 \ldots a_{k-1})}$

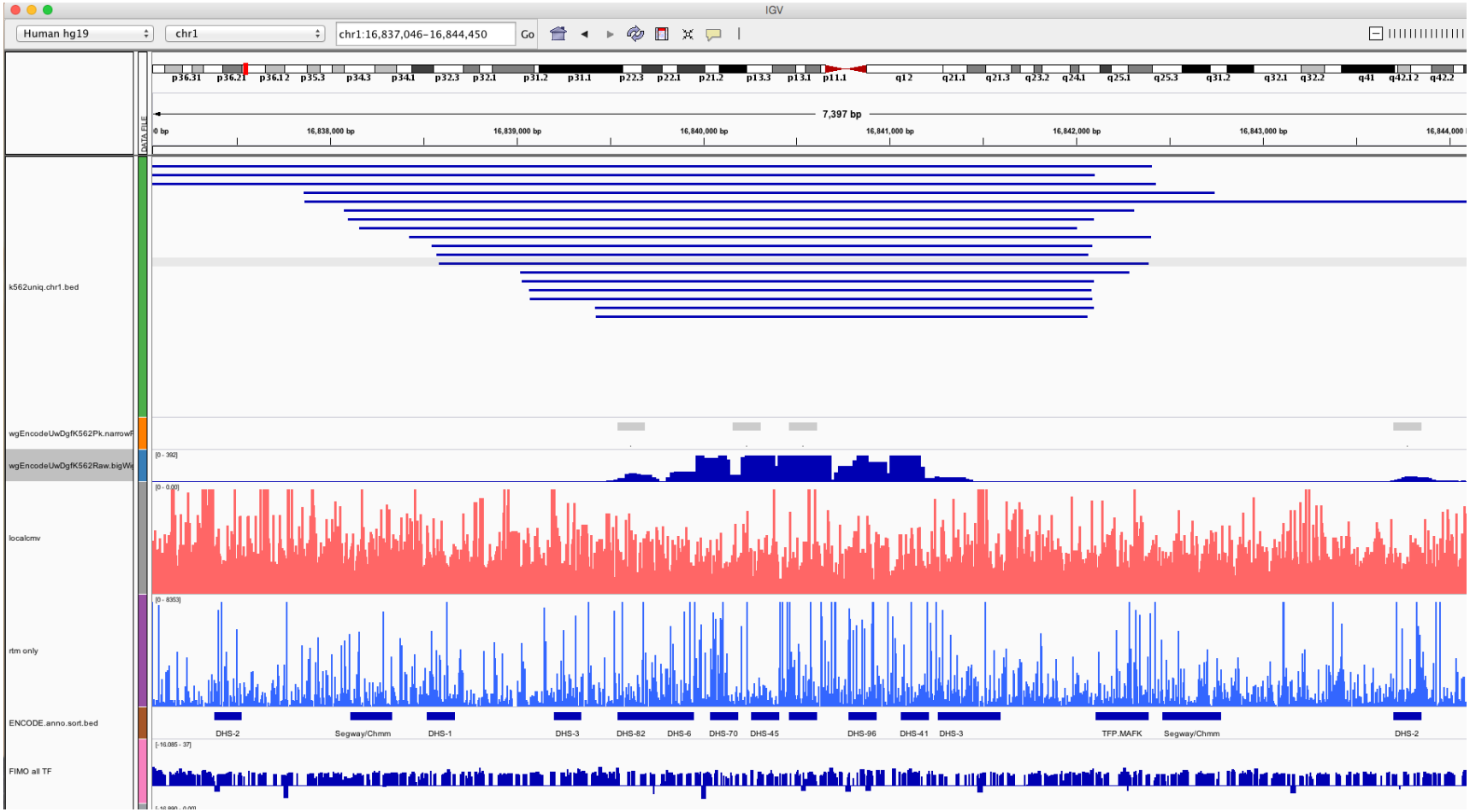
Return time distribution (RTD):
The shortest time(distance in bp, denoted by 'd') that to find the same kmer in the context of chromosome.

Define local reoccuring p-value:
$p(i) = (1 - p(k_i))^d \times p(k_i)$, $p(k_i)$ is the kmer composite frequency at position i

Yale

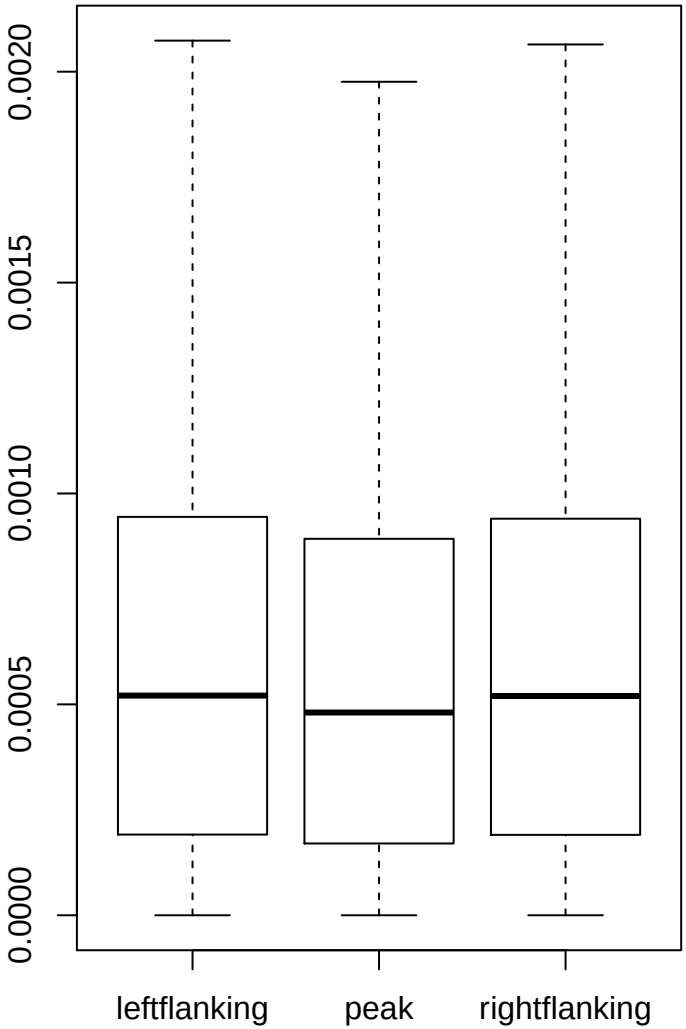
RTD adjusted local re-Occuring for binding peaks

whether there are specific pattern for peaks?



local-reoccuring

Compre the DNAse footprinting region with its 100bp far flanking region (on chr1):



Use K562 DNase footprintng peaks; and its left and right 100bp flanking
The kmer in the peaks has lower re-occurent rate.

Yale

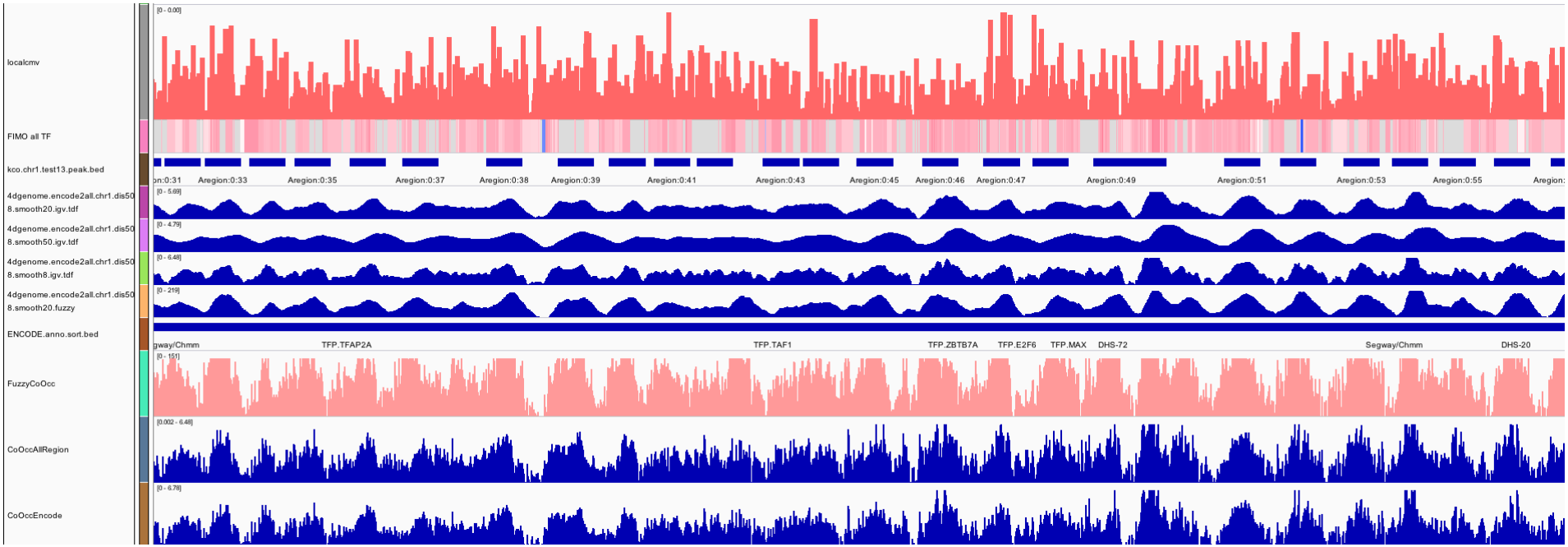
Kmer co-occurrence

The sequences in peak regions are highly associated than its flanking due the sequence specific binding (motif)
The kmer, say in the center of a peak, may have high co-occurence rate with its neighboring kmer in peak than kmers in flanking region.

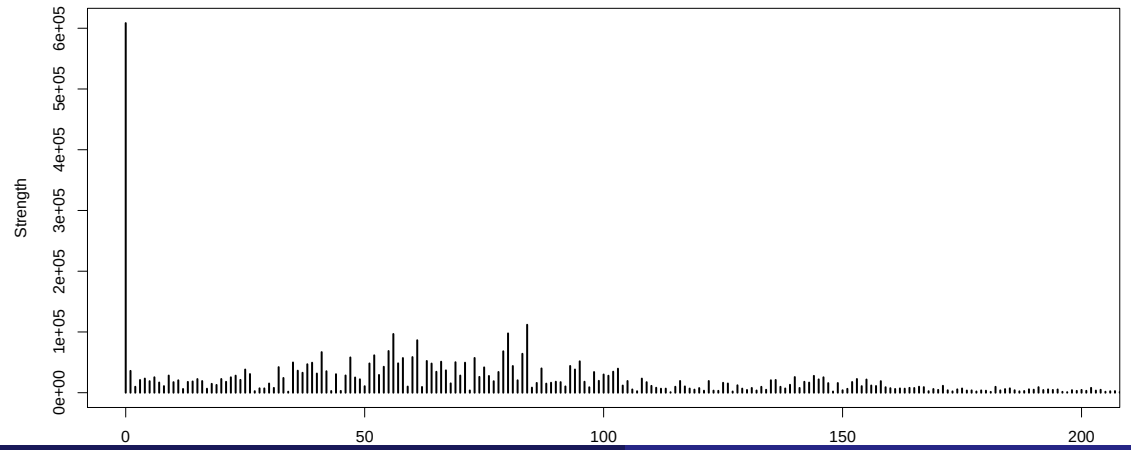
ch
Defined the kmer co-occurence as for $i \in \{1 \dots n\}$, k_i is the k-mer ended at the position i. $p(k_{peak}|k_i)$ and $p(k_{flanking}|k_i)$.
$K_{co} = \sum log(p(k_{peak}|k_i)) - \sum log(p(k_{flanking}|k_i))$

Yale

Evaluate these peaks

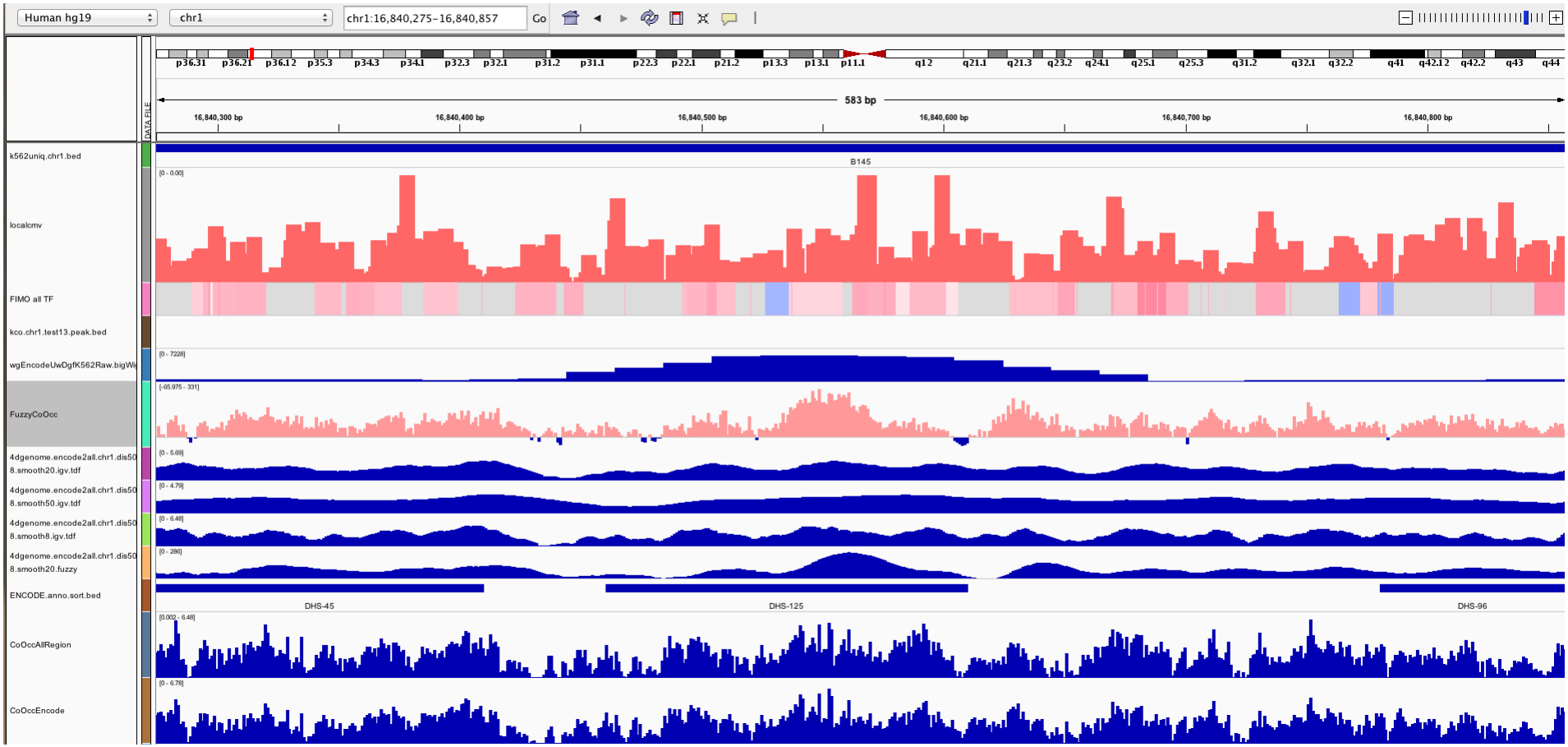


FFT

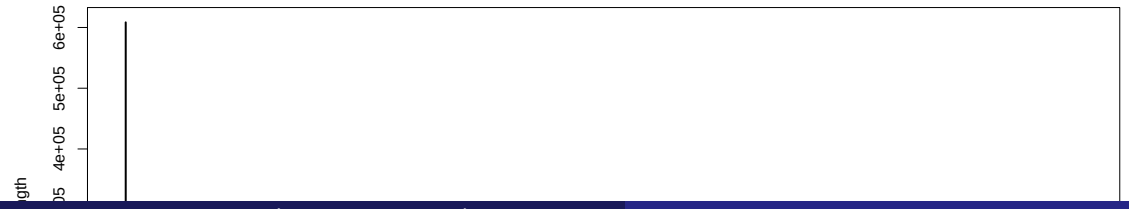


Yale

Evaluate these peaks

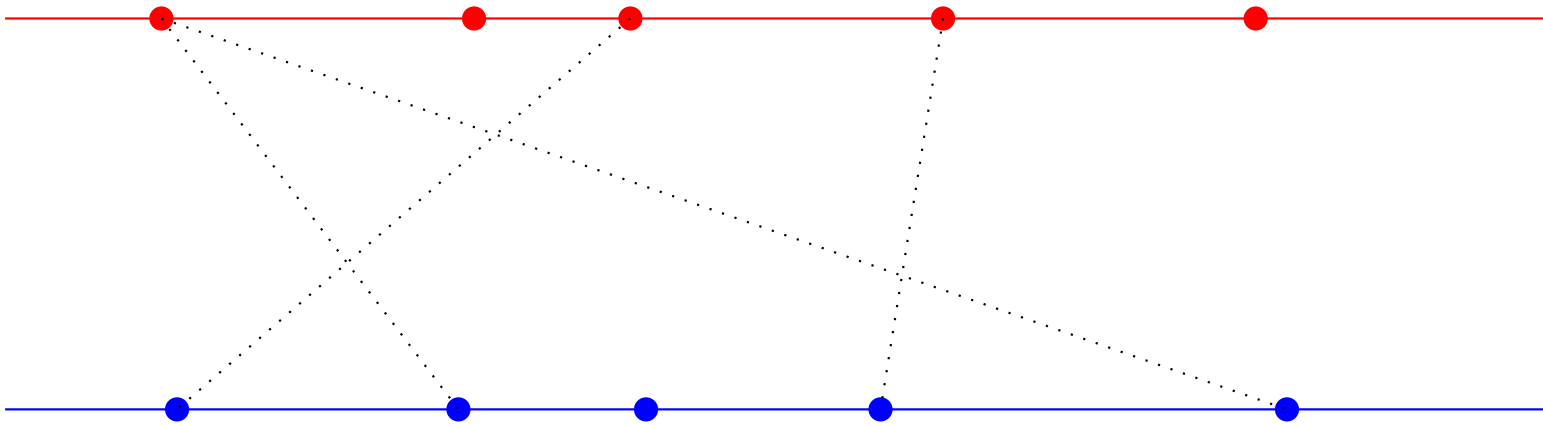


FFT



Co-occurrent for Peaks in AB region

A region



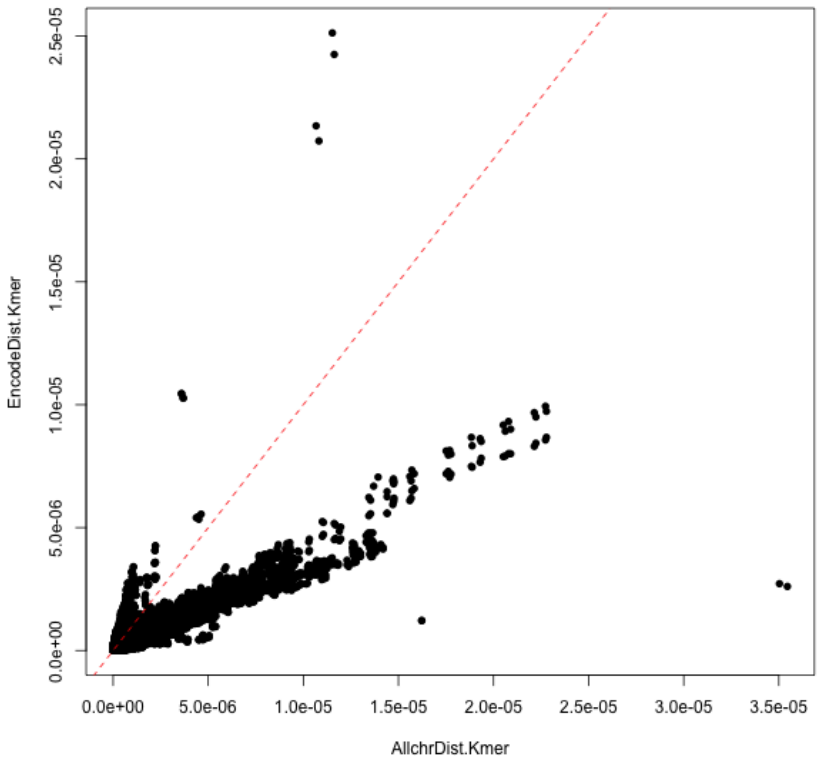
B region

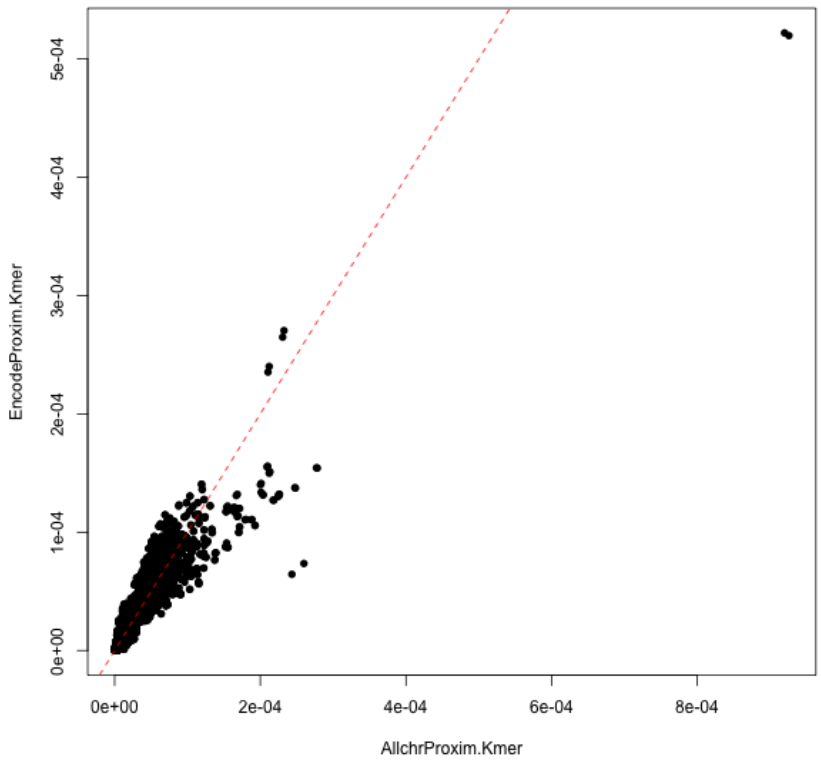
We hypothesis, the interaction between paired peaks from A,B, kmers from these peaks might have high chance to present in a close region(30bp-50bp).
For each k-mer in peak from A and B (peakA, peakB), I calculate the coccurence ratio between ENCODE peaks versus whole-genome.

Yale

Co-occurrence detection

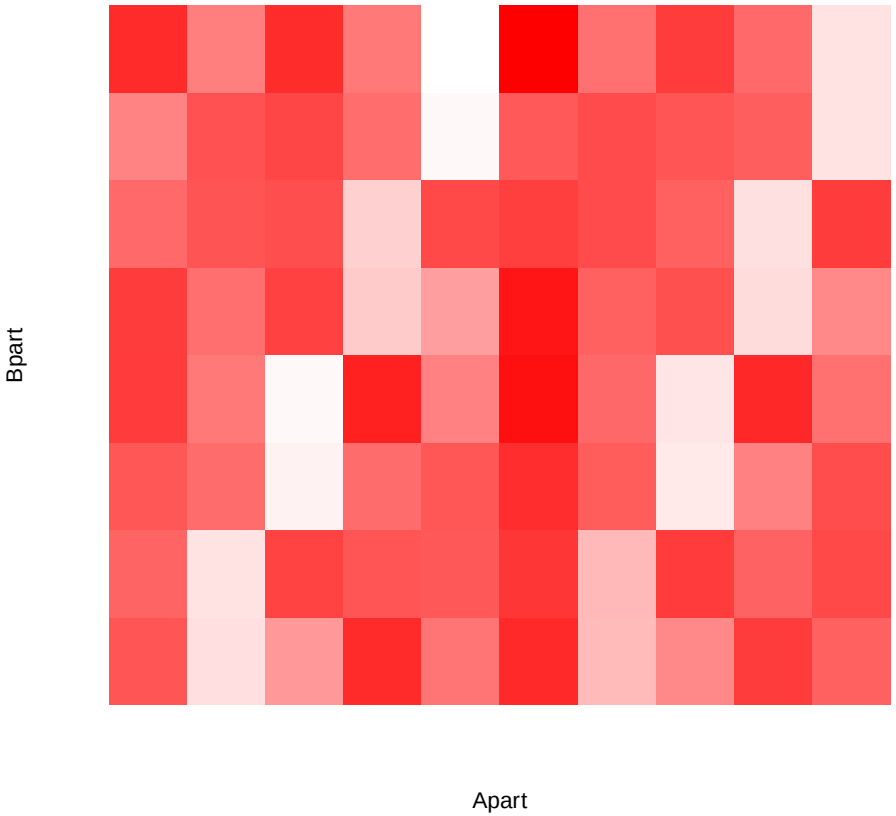
Use All encode enhancer peaks (from FunSeq2) and whole genome sequence, we calculated proximal(-18bp,+18bp] and distal [-50, -32] bp and [32,50]bp co-occurrence frequence and compare the differences.

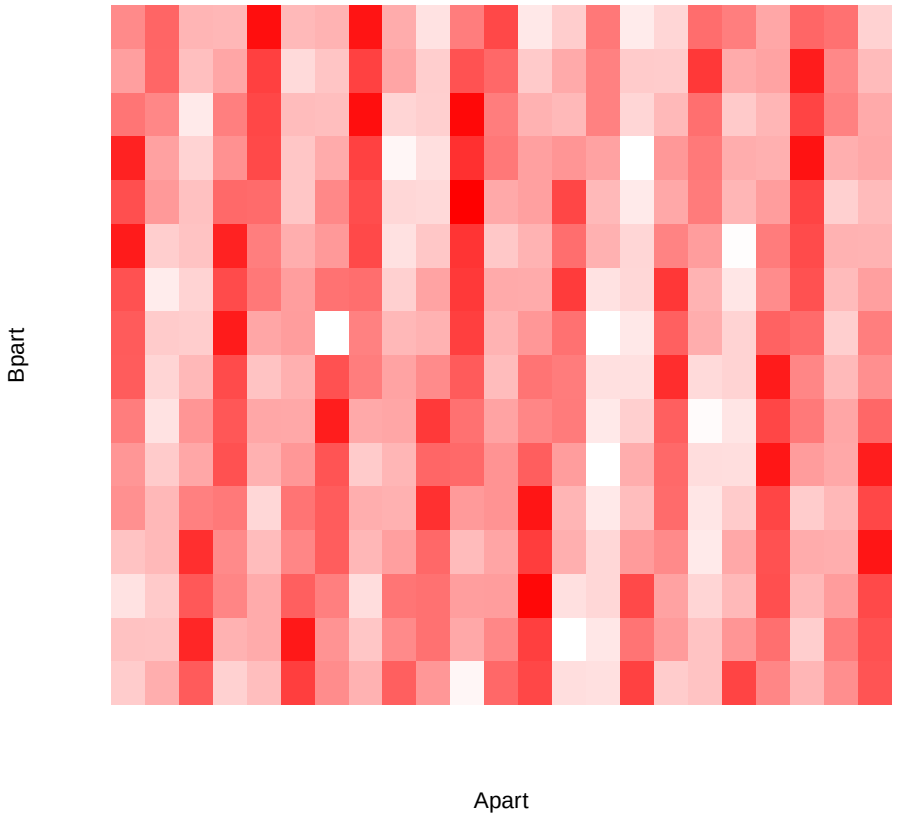




Yale

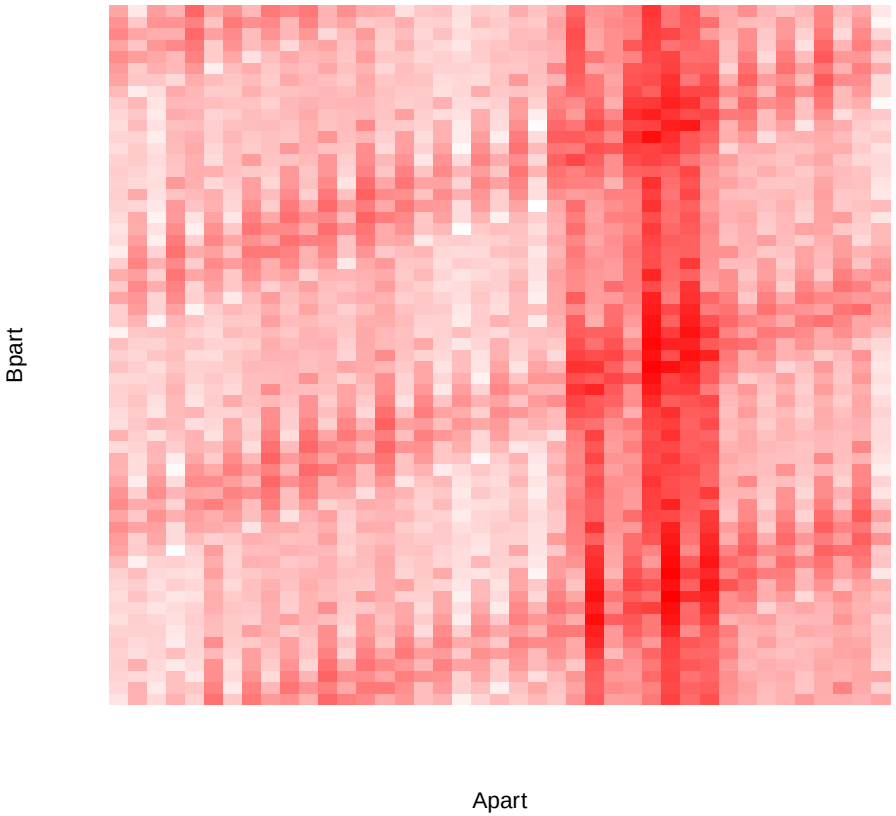
Co-occurrence Matrix

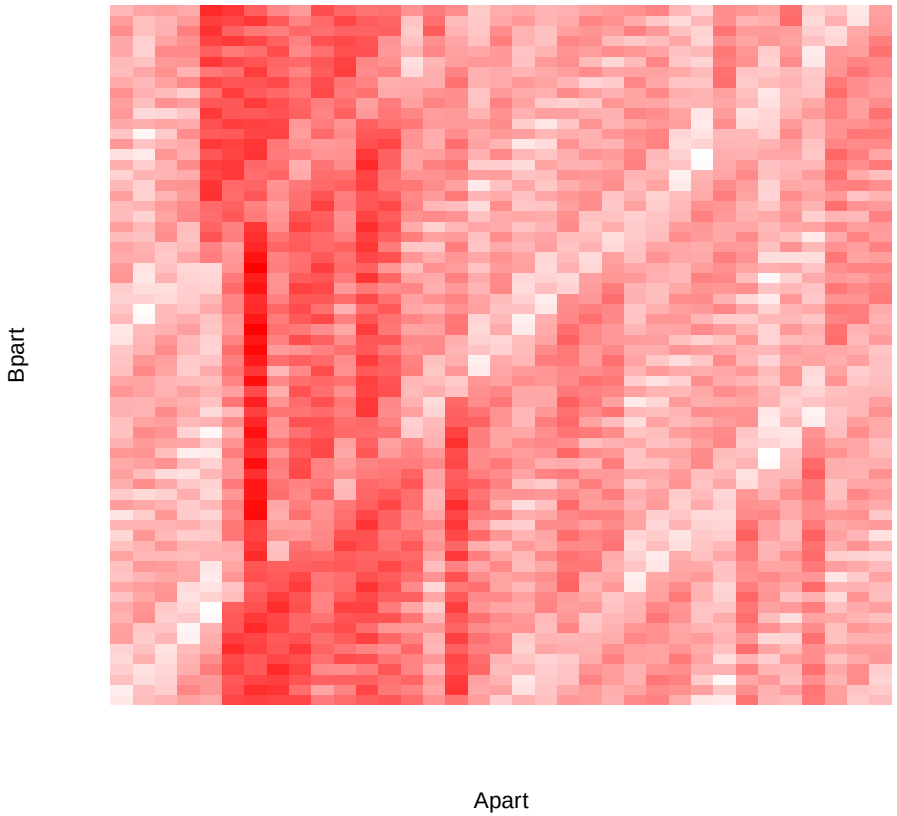




Yale

Co-occurrence Matrix





Yale

Known issues and solutions

- limitation of Kmer, exact matching
- affected by low-informative kmer, like AAAAAA
- doesn't employ positive set
- Whether hypothesis is correct?
- no spatial information considered

Solutions:

1. use known motif/ChIP-Seq peaks/FIMO for positive set
2. use kmer with/out specific gap (like 11011 pattern)
3. borrow idea from kmer-svm and gappedKmer(svm)

Yale

Future work

1. optimization of this framework
2. Define new way to evaluate
3. Expand it other chromatin signature, SNV

Yale