ENGINE: an Enhancer-Gene Interaction dEtection method using robust feature extraction

Lou Shaoke

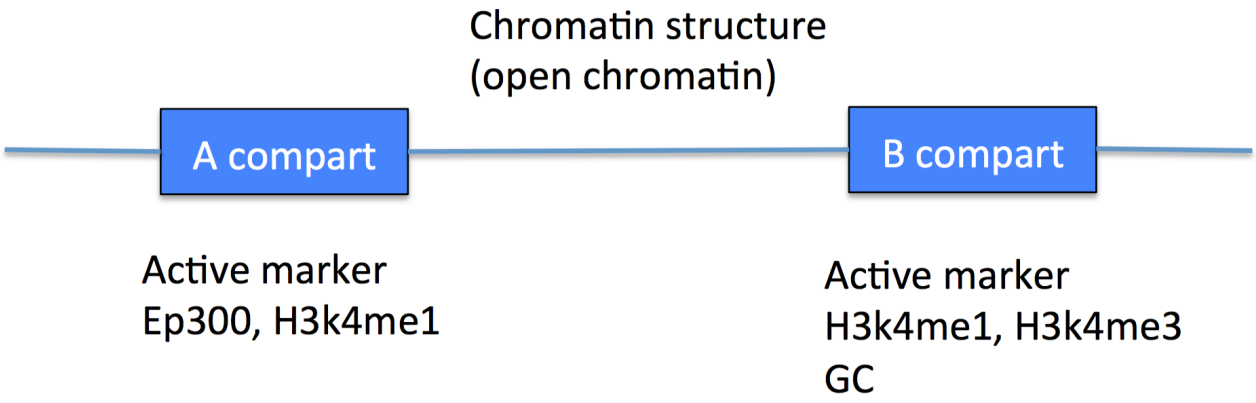
Department of Molecular Biophysics and Biochemistry

loushaoke@gmail.com

October 15, 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

Difficulties

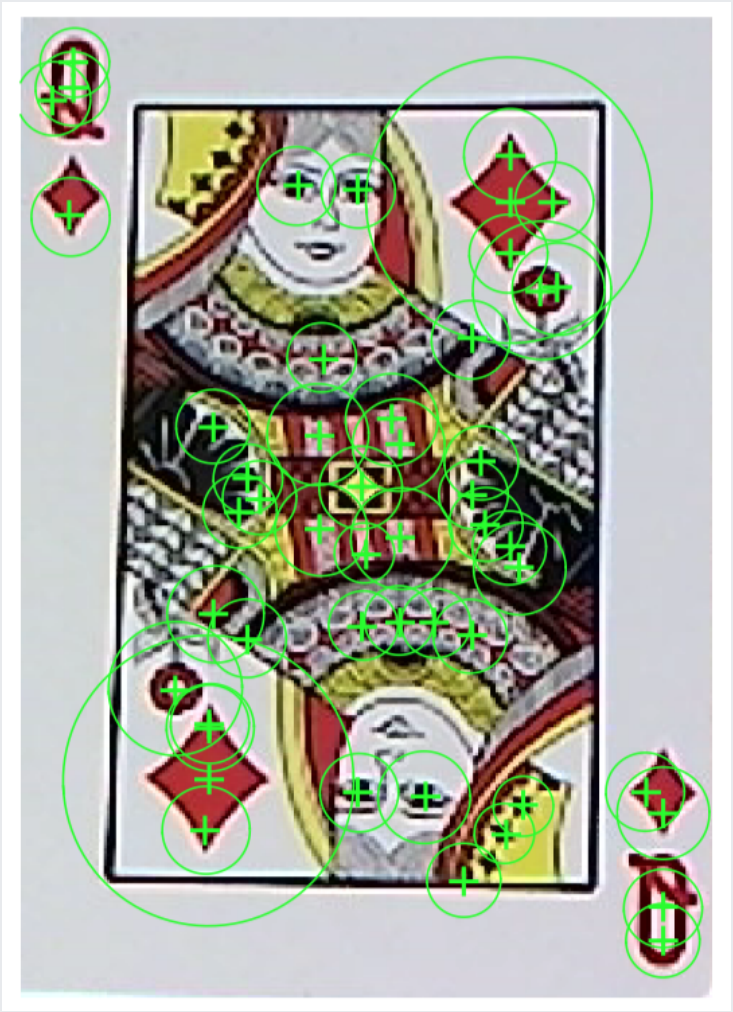
Given we have histone mark and other TF binding signal, it still has difficulties:

- Size of interaction region varies, Known methods: segementation (chromHMM and segway)
- No information of interaction site
- Two regions both have specific regulatory signal, No good methods to extract features.

Is it possible to identify robust features based on signals on A and B regions?

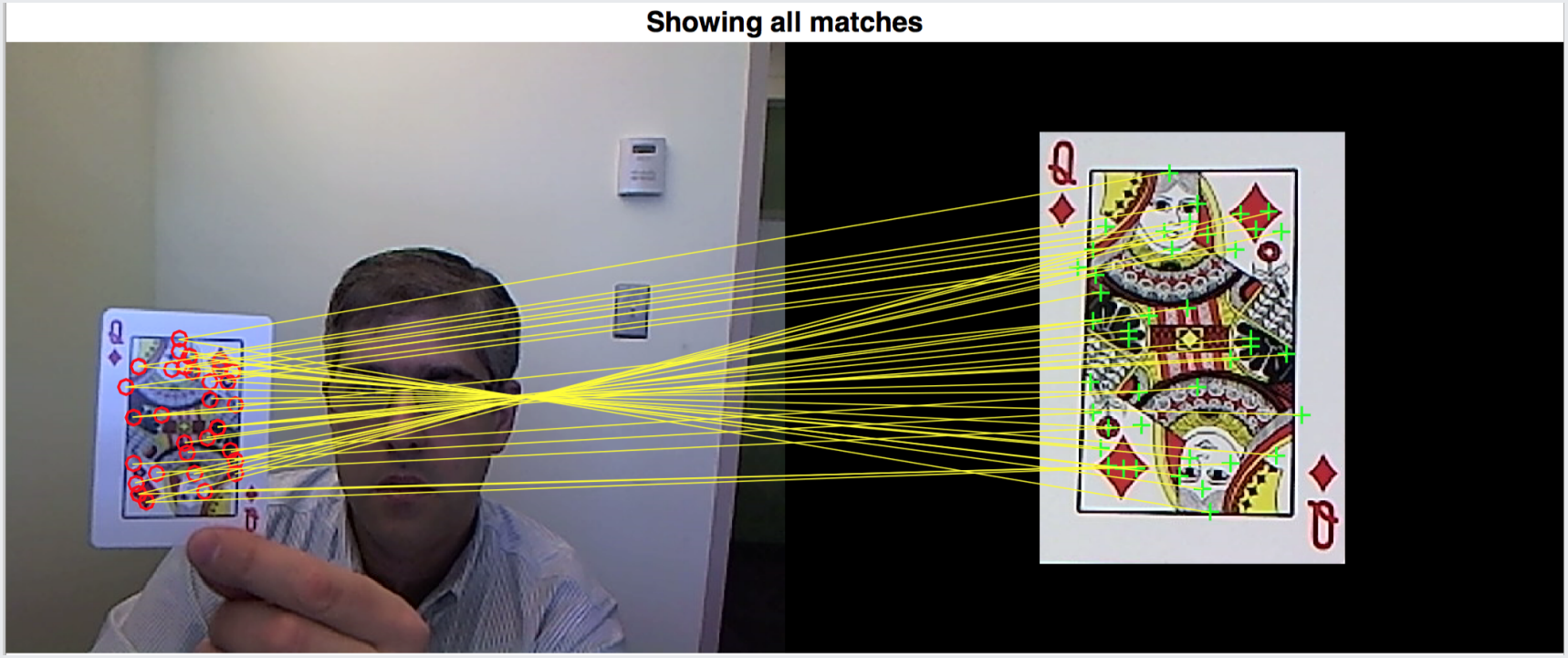
Feature extraction and recognation

Object recognition and pattern matching are popular in Computer vision.



Feature extraction and recognation

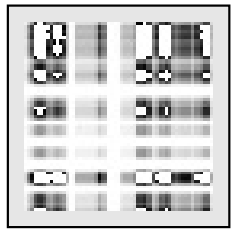
Object recognition and pattern matching are popular in Computer vision.



Flowchart

Flowchart

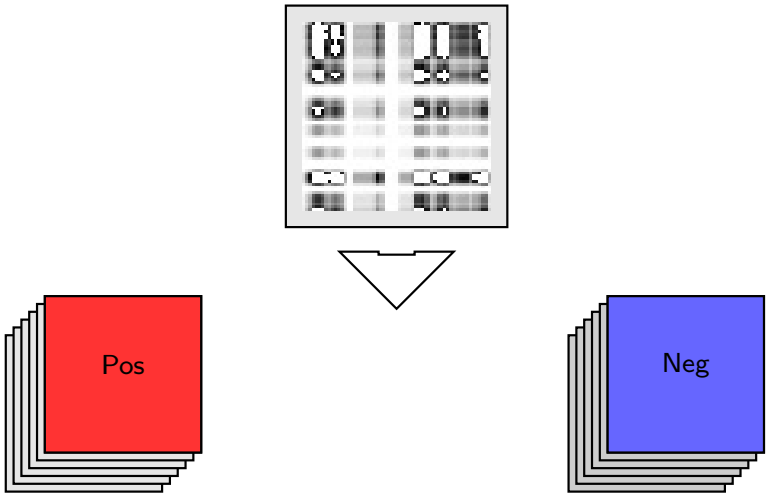
408 positive and 408 negative dataset, data transformation



Flowchart

SURF: Speeded Up Robust Features, merits:

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

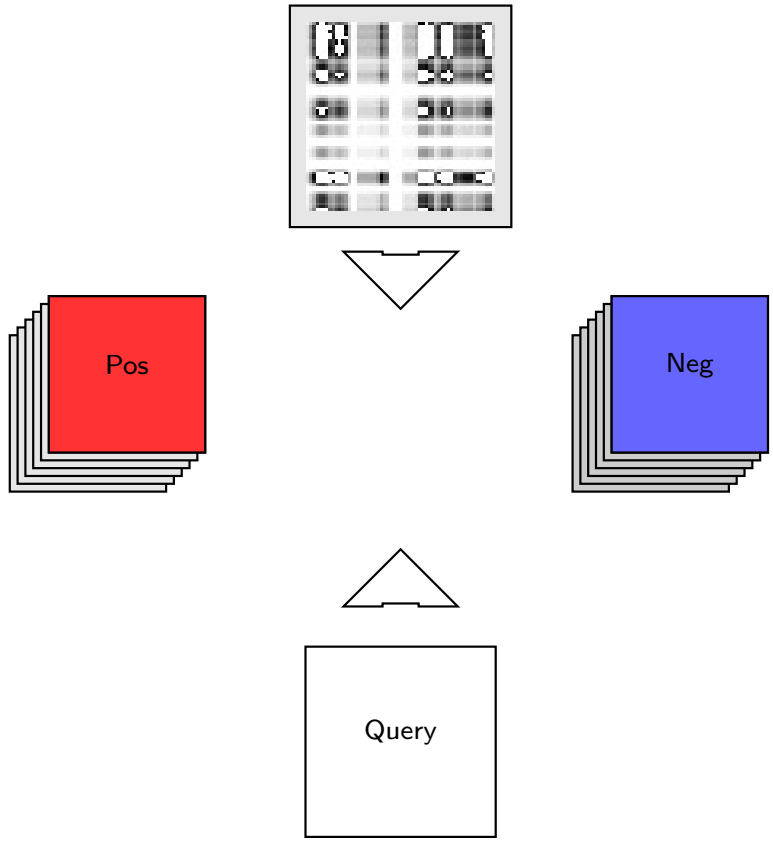


Flowchart

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

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

The enrichment score:
$ES(i) = -\sum log(\frac{\sum_j R_{i,j}}{N}) - log(\frac{\sum_j \sum_k 1\{s_i = n_j\}}{\sum_j \sum_k 1})$.
The relative enrichment score
$RS = ES(positive) - ES(negative)$.
The lower of RS, the better!



Flowchart

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

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

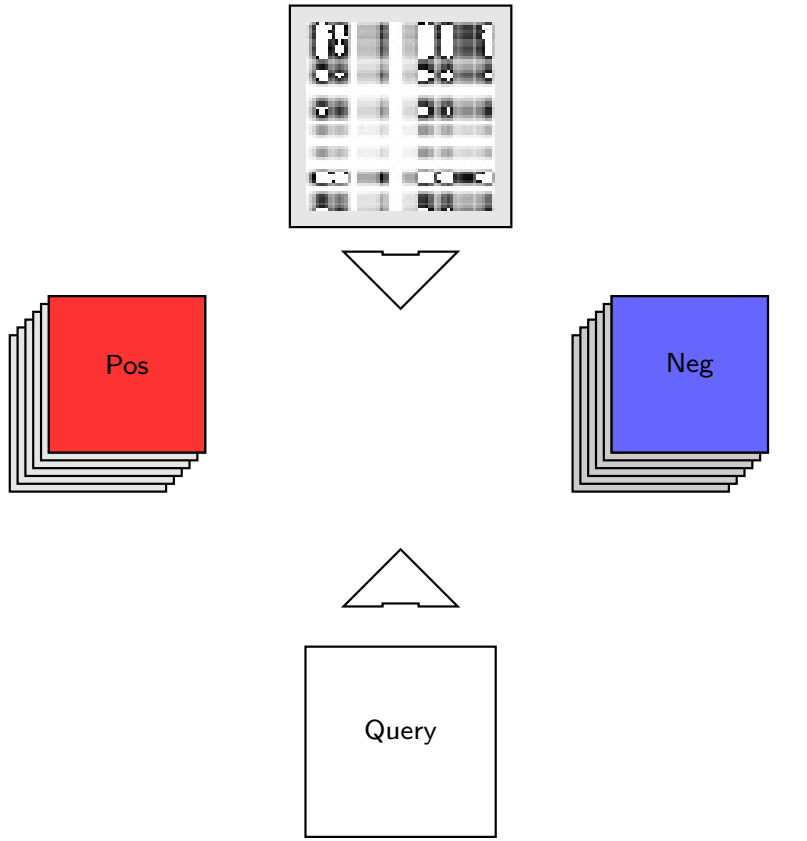
The enrichment score:

$$ES(i) = -\sum log(\frac{\sum_j R_{i,j}}{N}) - log(\frac{\sum_j \sum_k 1\{s_i = n_j\}}{\sum_j \sum_k 1}).$$

The relative enrichment score

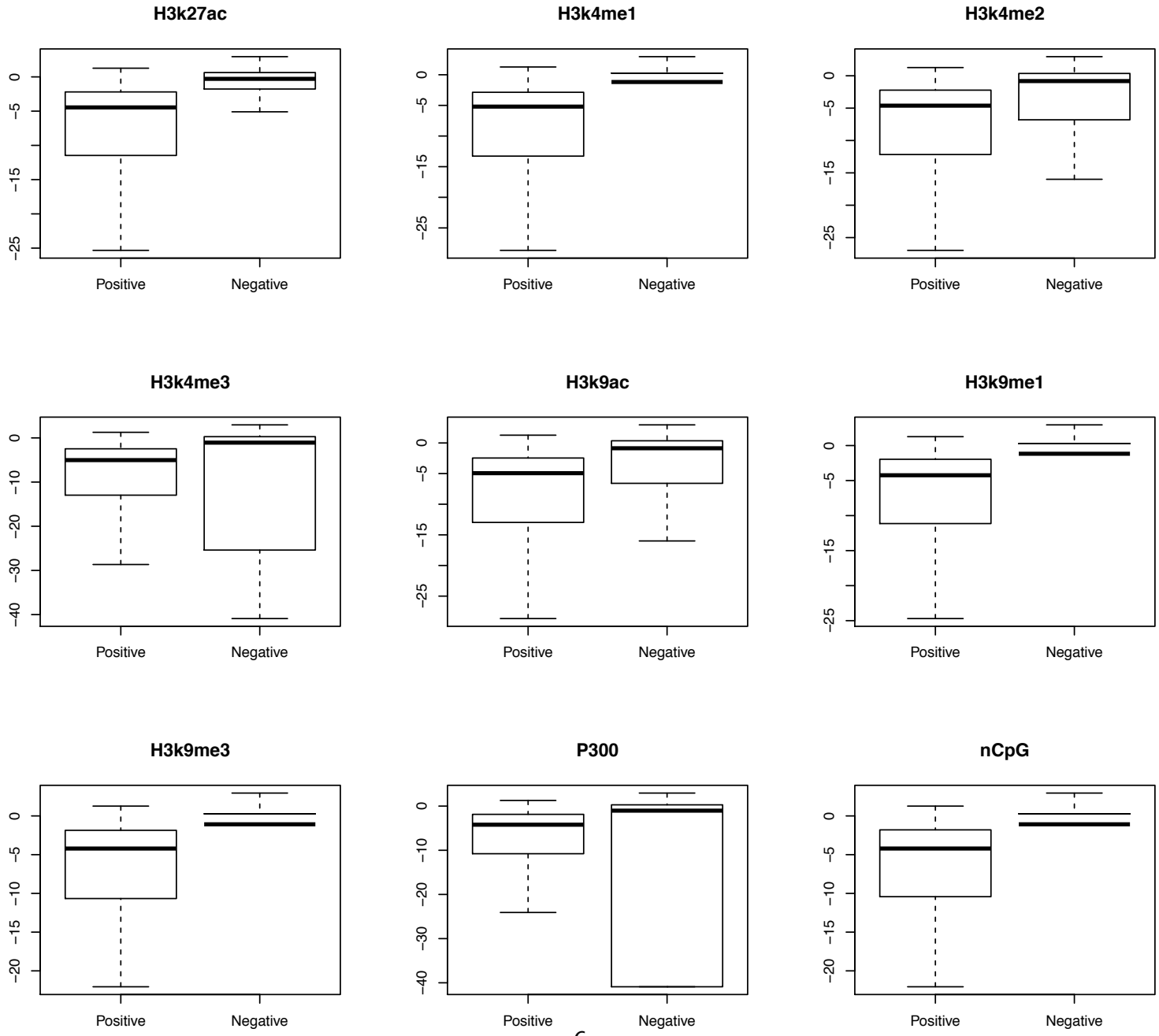
$$RS = ES(positive) - ES(negative).$$

The lower of RS, the better!



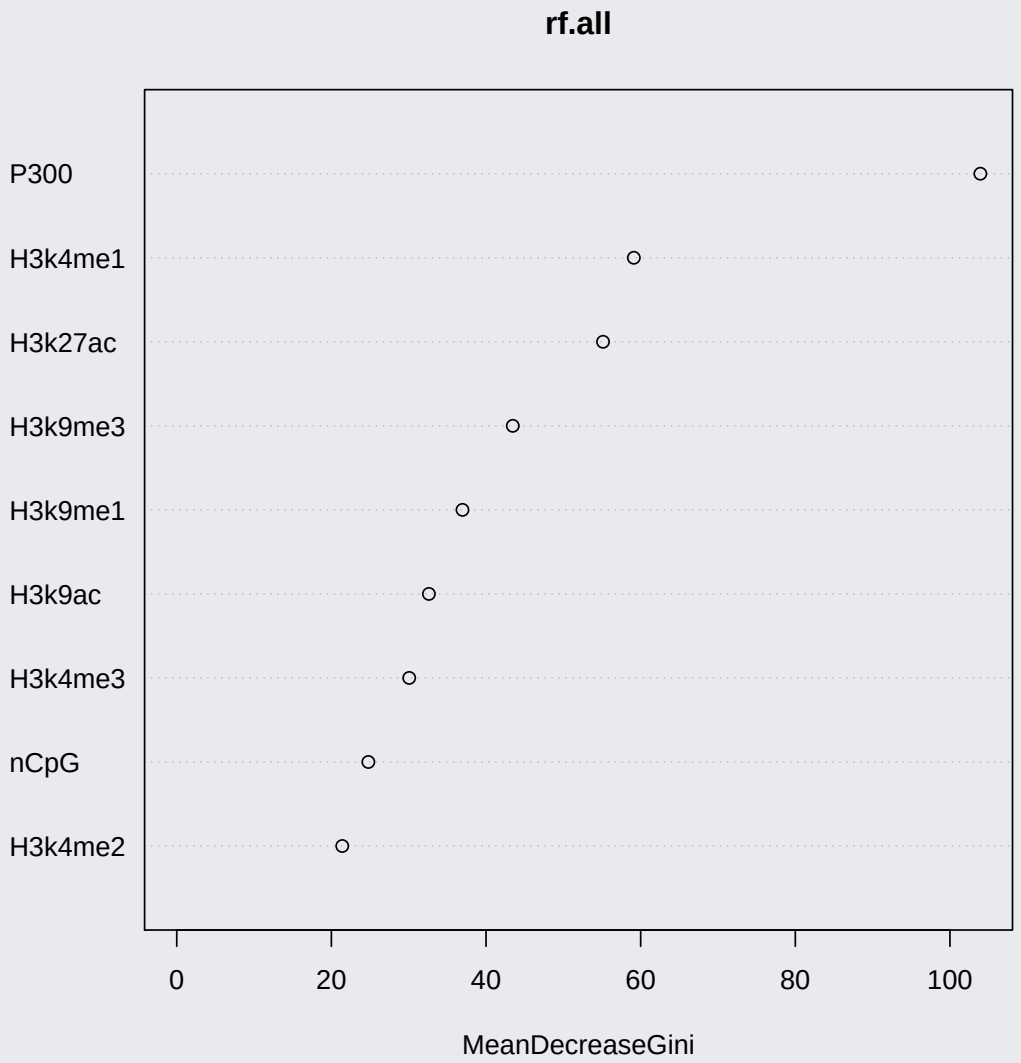
H3k27ac,H3k4me1,H3k4me2,H3k4me3,H3k9ac, H3k9me1,H3k9me3,P300

RS distribution:



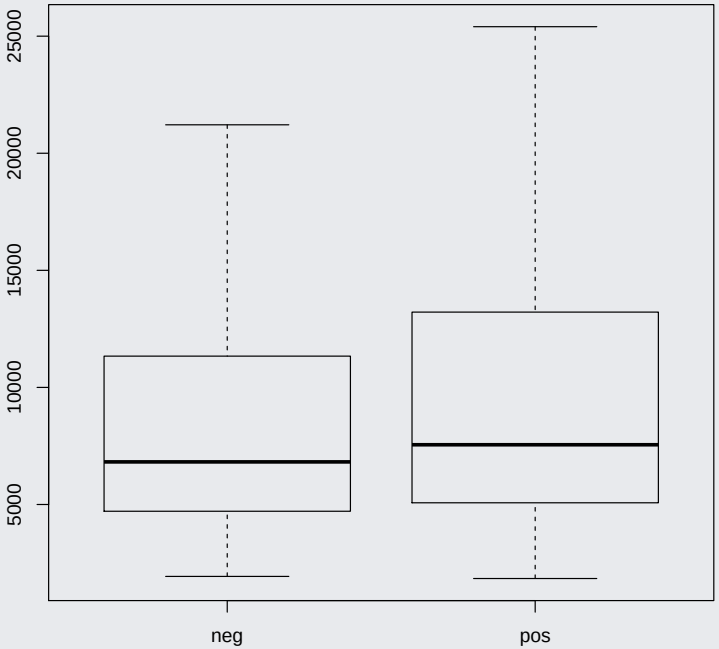
RandomForest

AUC:0.958, feature importance:



Discussion

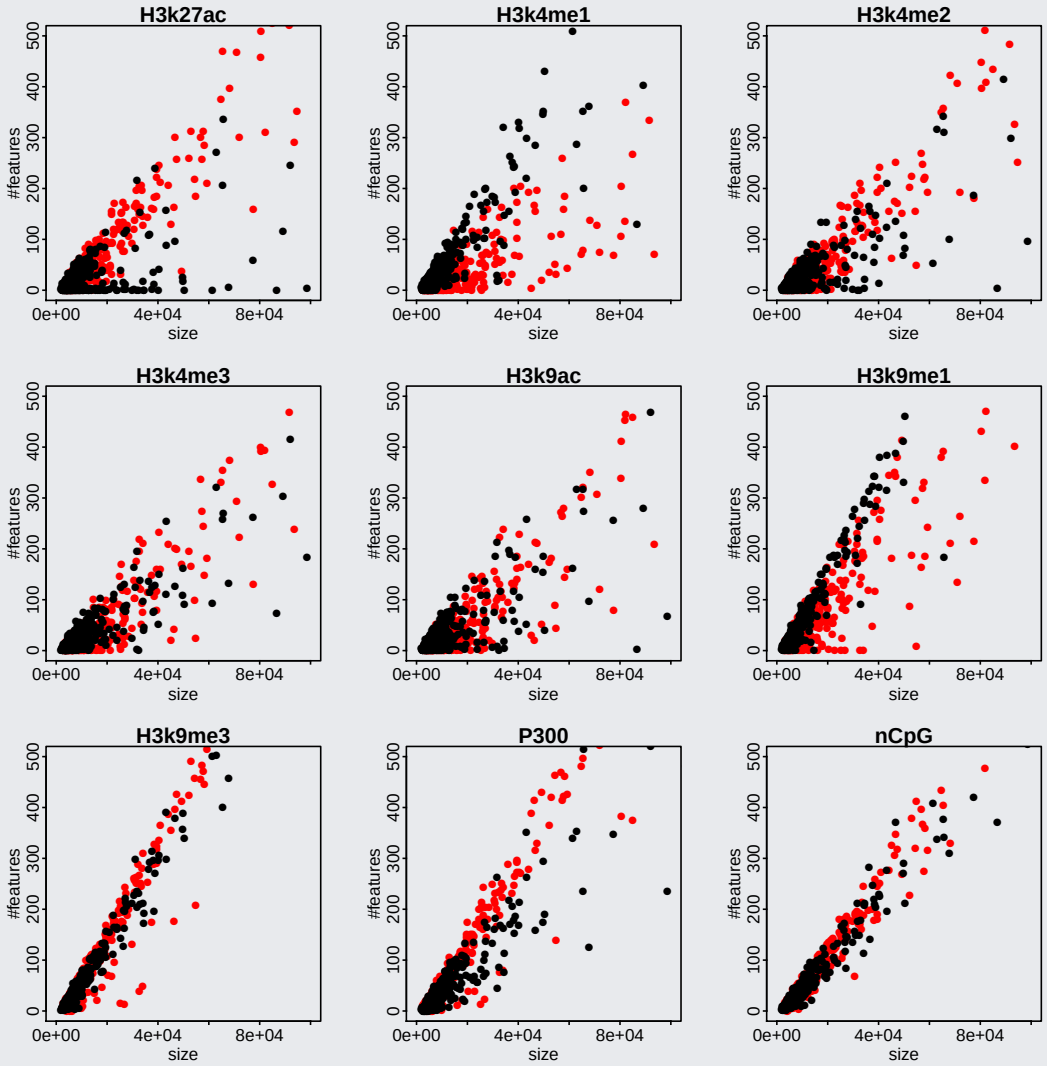
- Too good to be true? size bias? feature extraction black box?



- computational intensity
- Biological meaningful? How to link the extracted features with functions?
- Integrate the previous Kmer co-occurence feature.
- Evaluation?

Supplementary

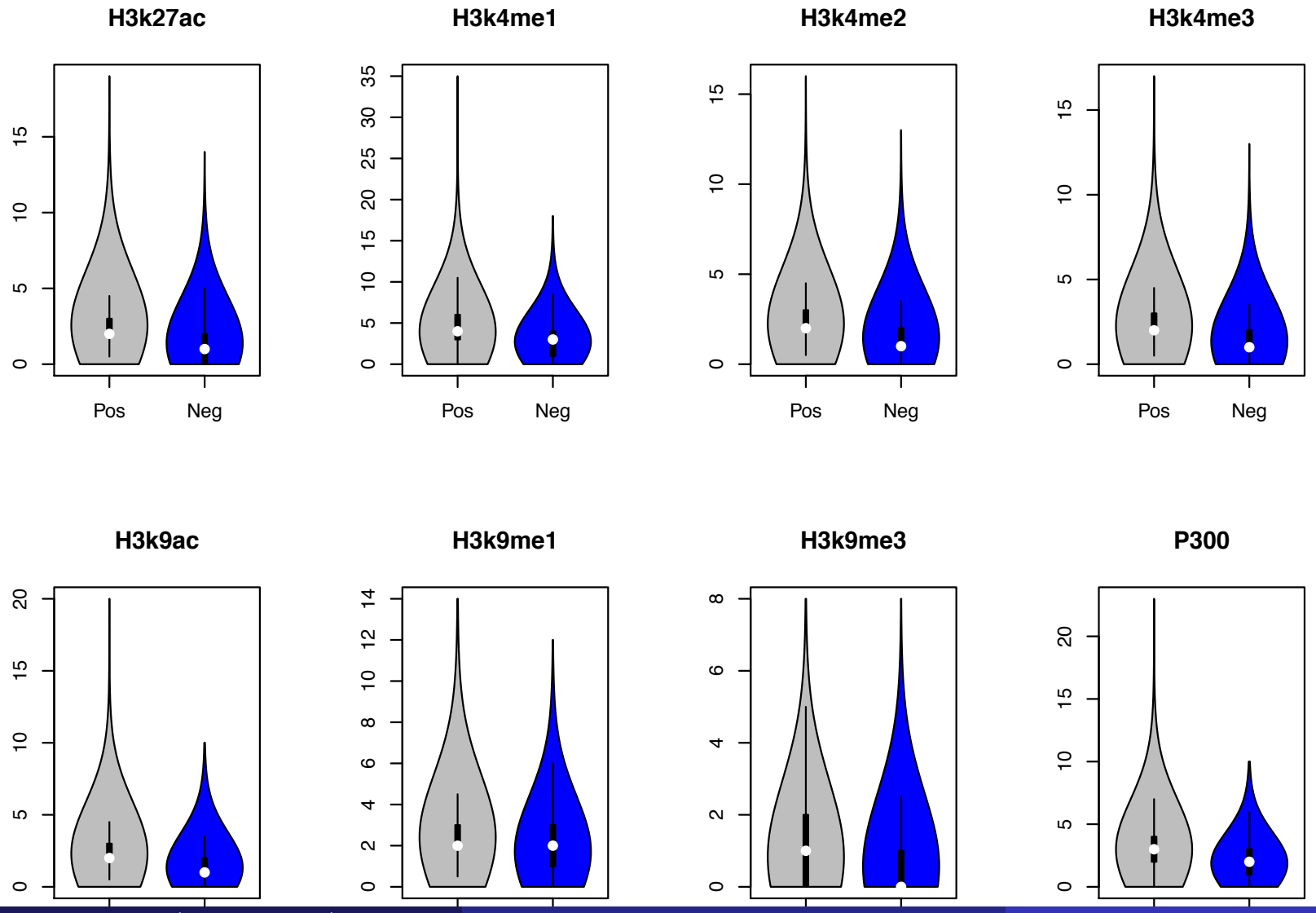
size vs #features:



positive set; negative set

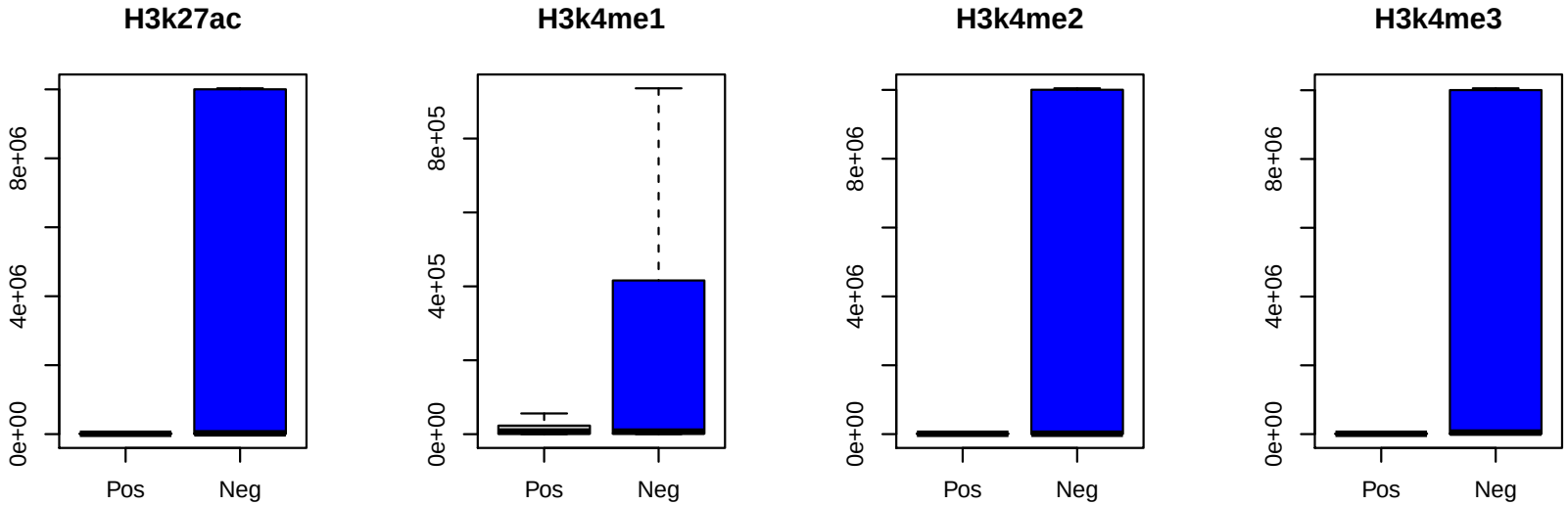
peaks

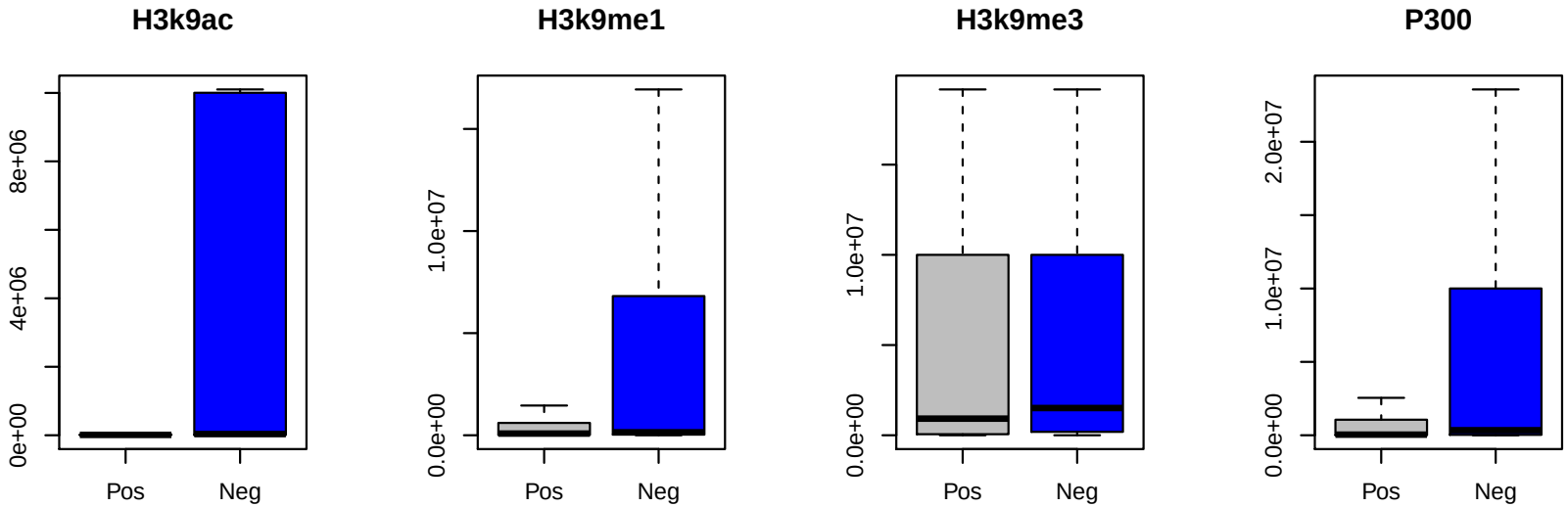
The number of histone mark peaks overall in positive region is more than that in negative region.



Peak size comparison

We also compared the smalled and biggest peaks in these around region.





Yale

Signal

Limitation of data: peaks called, especially with replicas, there are no uniform peaks calling peaks, which results in the overlap of peak region.
Using the signal, but we need to face:
No good way to segment the 3d interaction region into features: no fixed-window, varied region size(sometime need.
idea is to convert the AB region feature into a image like feature matrix and then use computer vision method to extract features

Yale