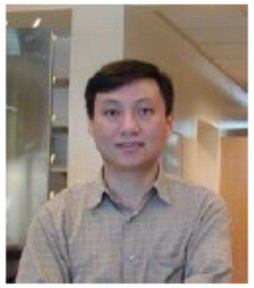
Hi-C updates

KKY

1. Reproducibility and QC metrics in ENCODE 3D nucleome subgroup
2. Identifying topologically associating domains in multiple resolutions

Updates of the ENCODE 3D nucleome subgroup

- Preparation of manuscript for ENCODE guidelines for assessing the quality and the reproducibility of chromosome conformation capture experiments
    - Similar to ENCODE ChIP-seq guidelines (Landt et al. Genome Research 2012)

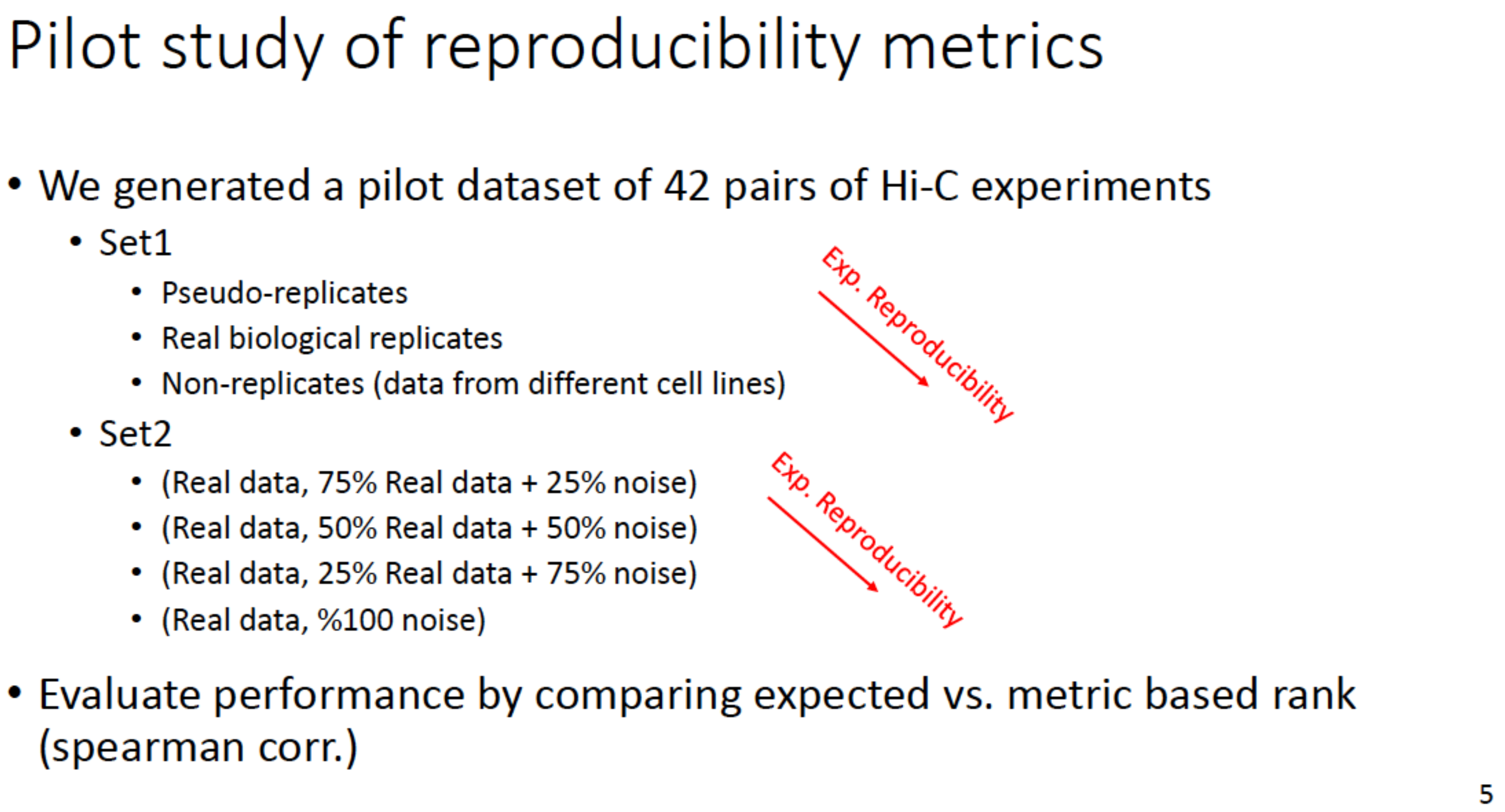


Hi-C data
11 cell types
2 **replicates**

Hi-C data
Mouse
forebrain
Time course
2 **replicates**

slide adapted from Noble lab

Evaluate reproducibility metric



Pilot study of reproducibility metrics

- We generated a pilot dataset of 42 pairs of Hi-C experiments
  - Set1
    - Pseudo-replicates
    - Real biological replicates
    - Non-replicates (data from different cell lines)
  - Set2
    - (Real data, 75% Real data + 25% noise)
    - (Real data, 50% Real data + 50% noise)
    - (Real data, 25% Real data + 75% noise)
    - (Real data, %100 noise)
- Evaluate performance by comparing expected vs. metric based rank (spearman corr.)

Exp. Reproducibility

Exp. Reproducibility

5

slide adapted from Noble lab

Quantifying reproducibility using spectral graph theory

Laplacian $L = D - A$

A is the contact matrix
D is a diagonal matrix
such that $D_{ii} = \sum_j A_{ij}$

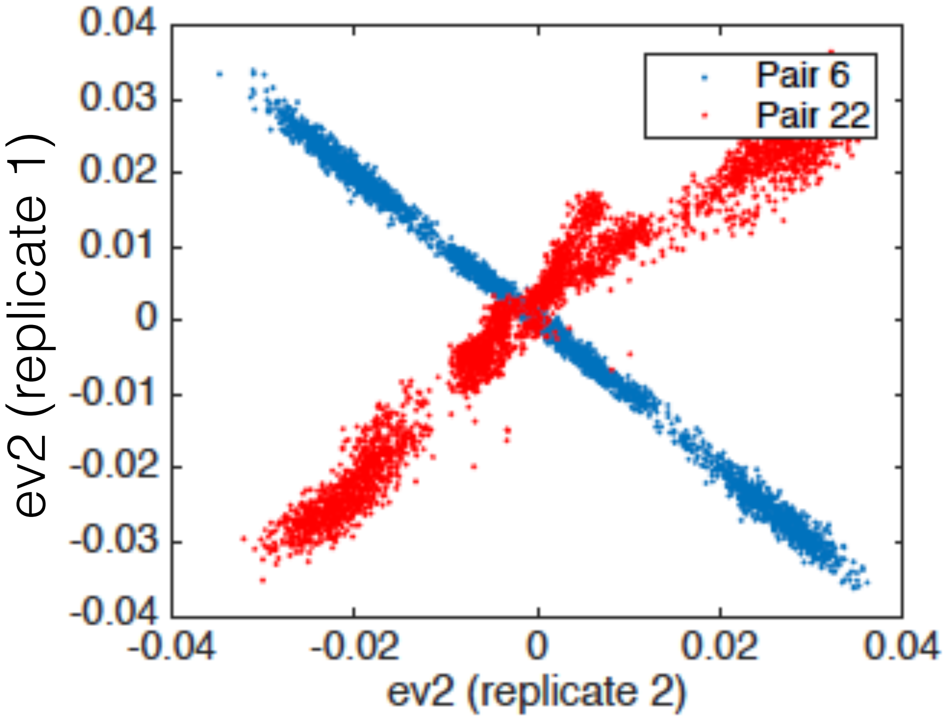
$$\mathcal{L} = I - D^{-1/2} A D^{-1/2}$$

$$0 = \lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \ldots \leq \lambda_n$$

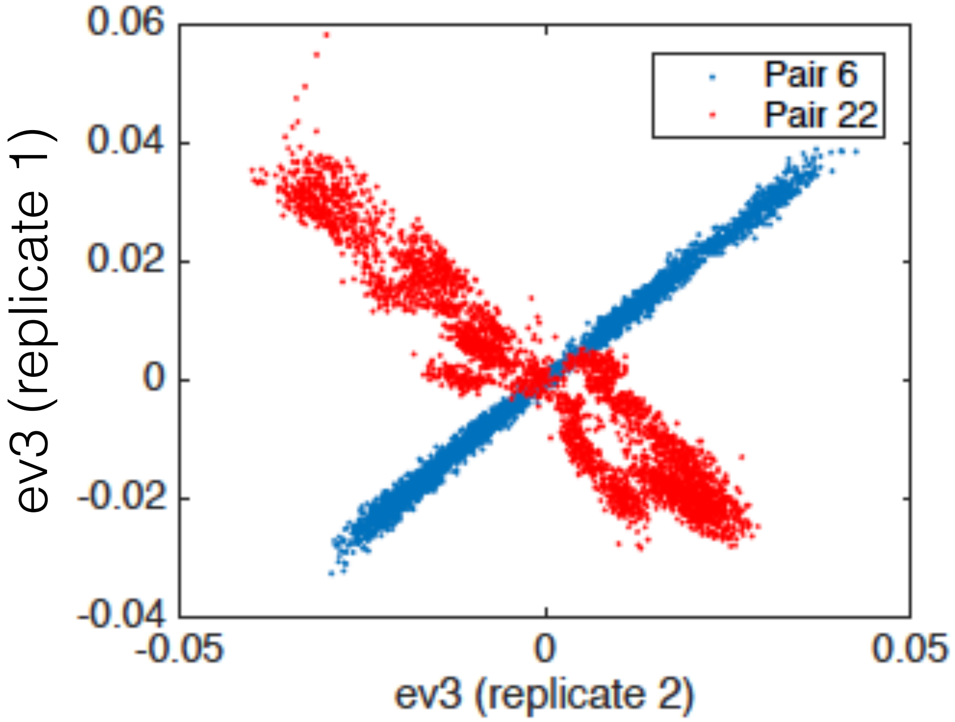
- leading eigenvectors capture the structures of the graph (dimension reduction)

	A	A’	Euclidean distance
leading 5 eigenvectors	ev1	ev1’	d1
	ev2	ev2’	d2
	ev3	ev3’	d3
	ev4	ev4’	d4
	ev5	ev5’	d5

score= sum over d

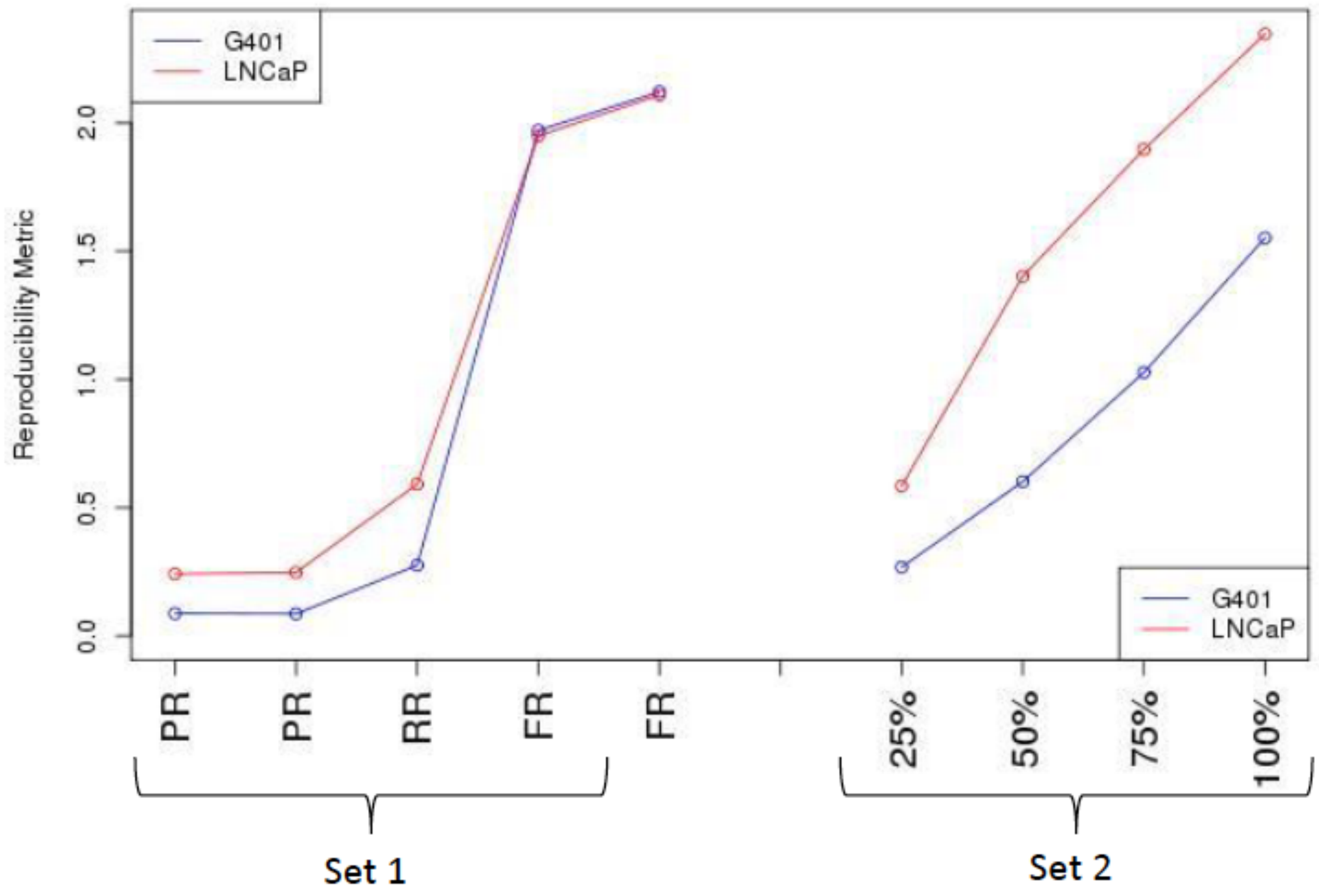


Pair 6 is more reproducible than pair 22

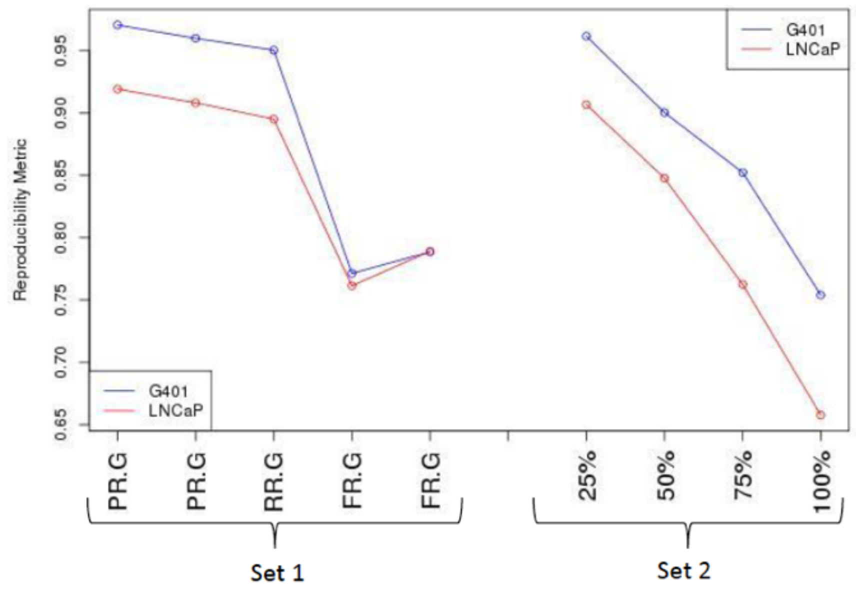


Results of various metrics

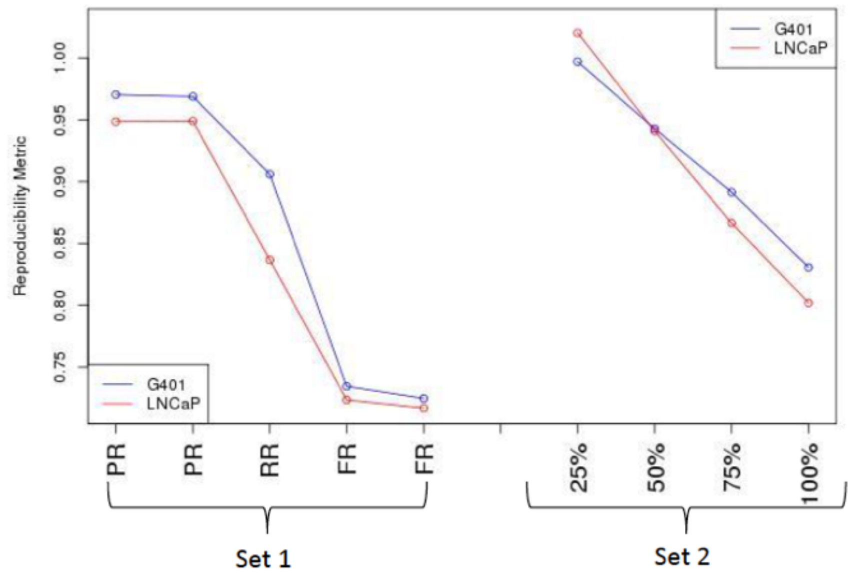
KKY



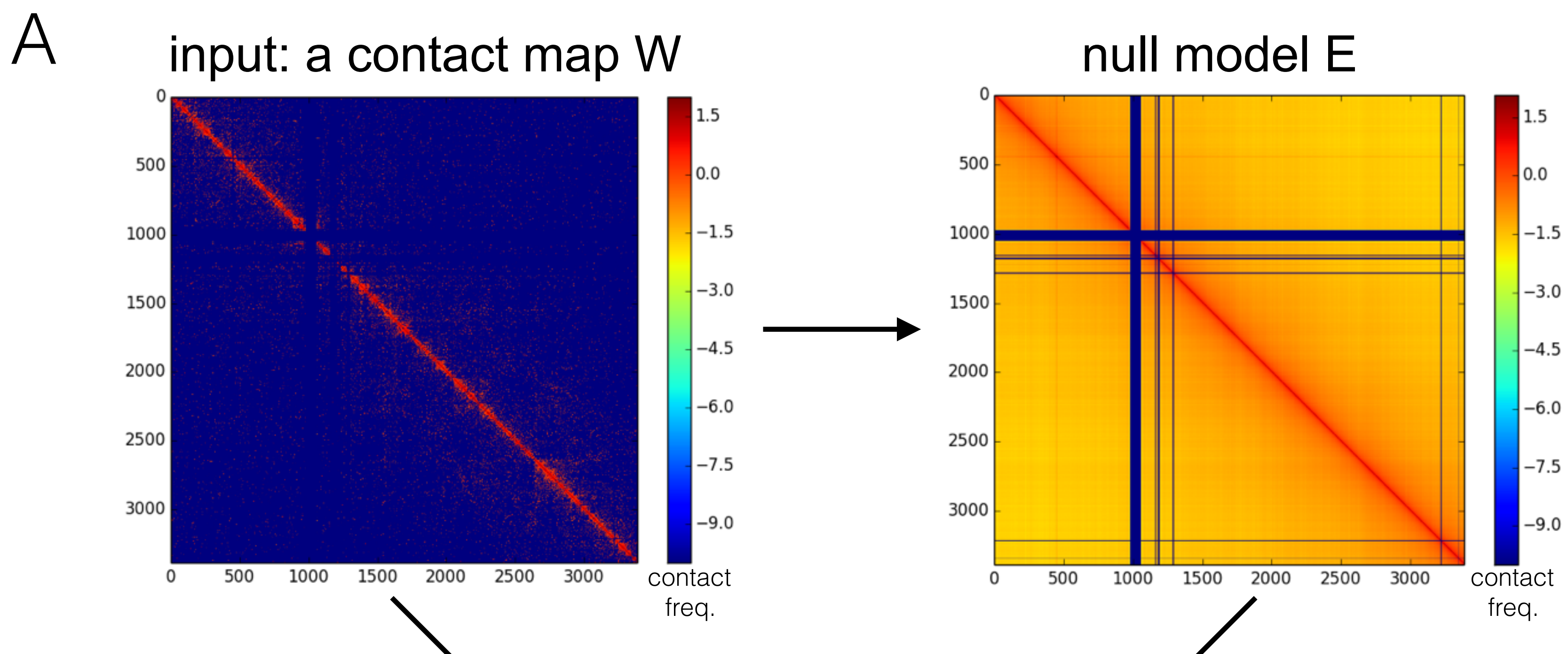
Li lab, Stratification-Aggregation



Kundaje lab, wavelet



- Studies of reproducibility and QC metric in ENCODE 3D nucleome subgroup
- Identifying Topologically associated domains in multiple resolutions

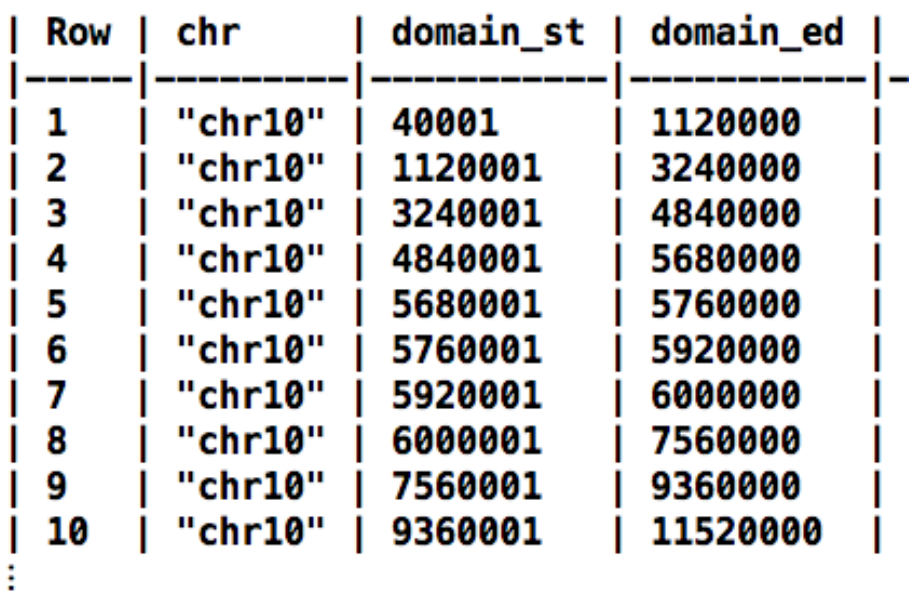


$$E_{ij} = c_i^* c_j^* f(|i-j|)$$

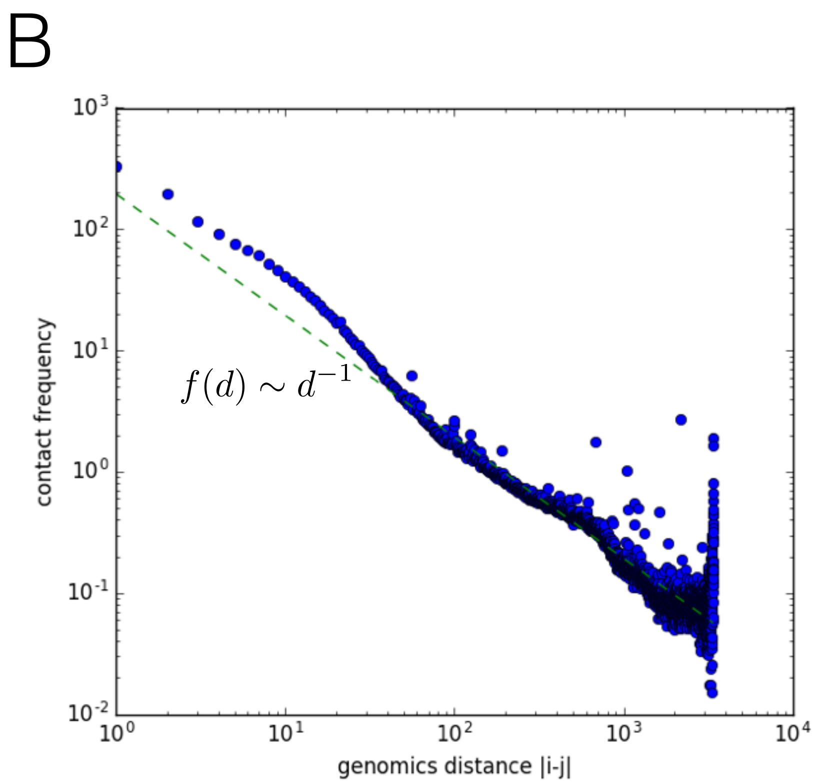
- choose a particular resolution
- optimize Q over all possible partitions
- multiple runs to increase robustness

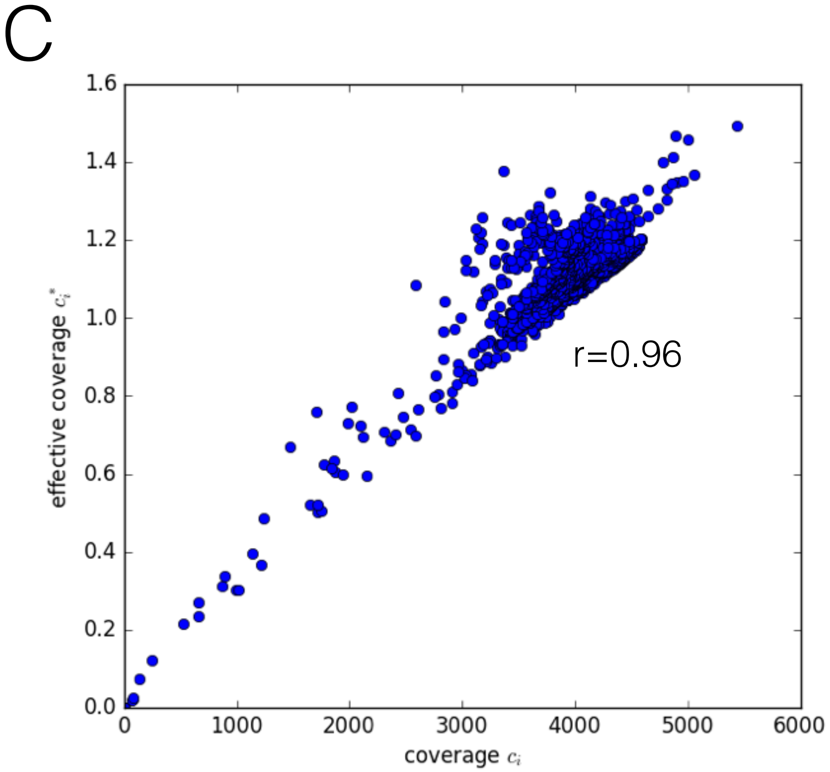
$$Q = \frac{1}{2N} \sum_{ij} (W_{ij} - \gamma E_{ij}) \delta_{\sigma_i \sigma_j}$$

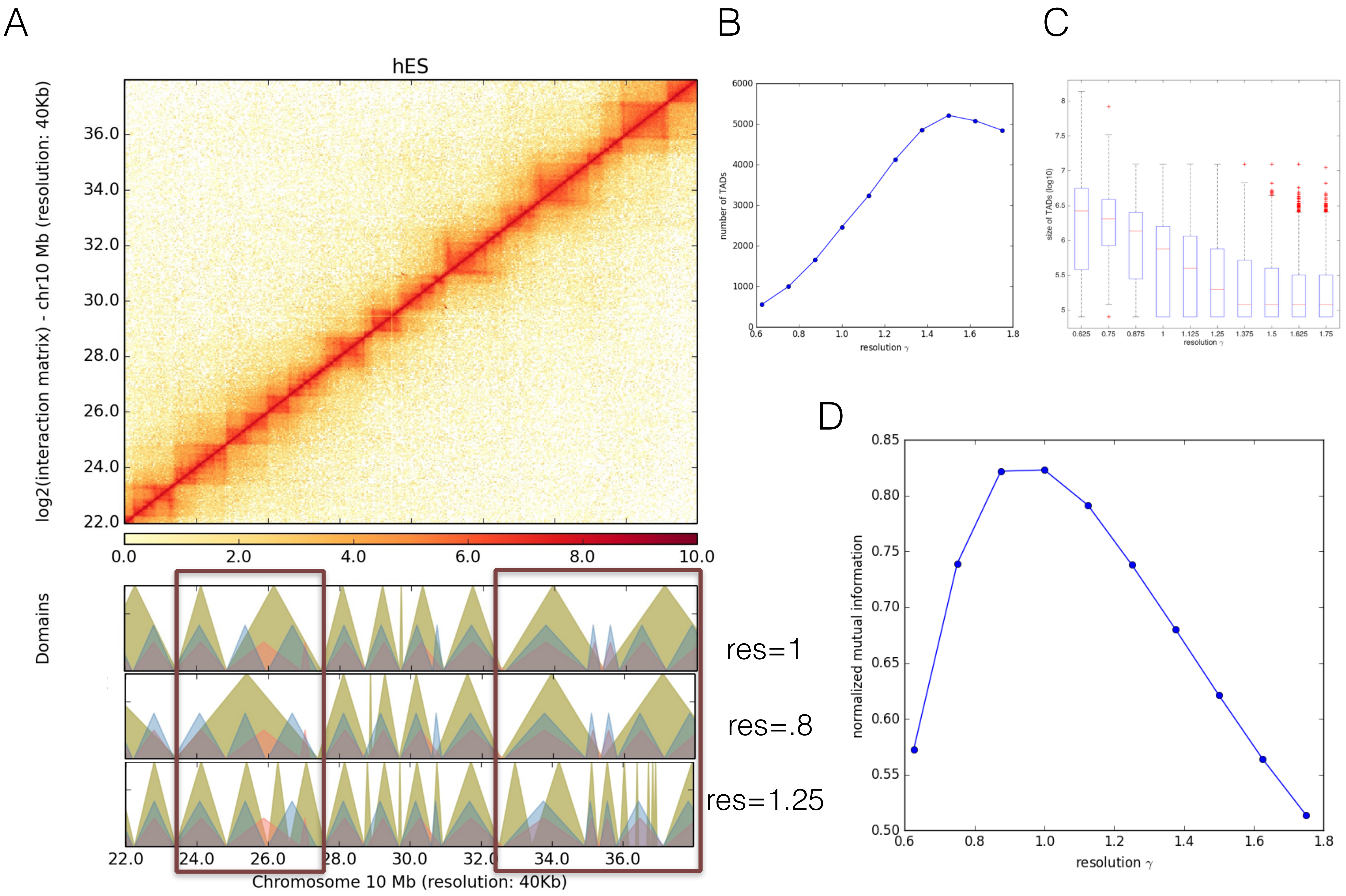
Output:



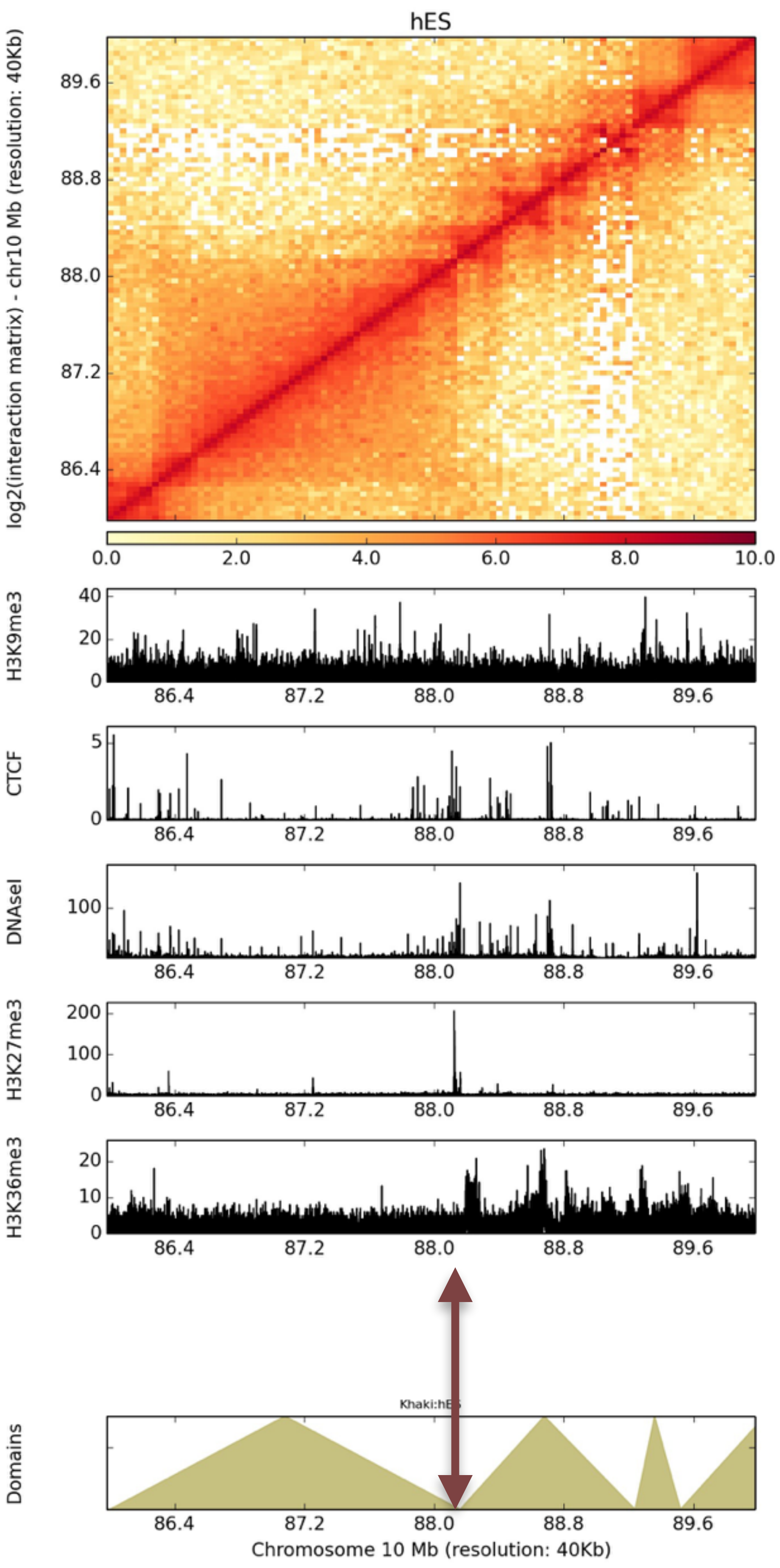
```
| Row | chr      | domain_st  | domain_ed  |
|-----|----------|------------|------------|-
| 1   | "chr10" | 40001      | 1120000    |
| 2   | "chr10" | 1120001    | 3240000    |
| 3   | "chr10" | 3240001    | 4840000    |
| 4   | "chr10" | 4840001    | 5680000    |
| 5   | "chr10" | 5680001    | 5760000    |
| 6   | "chr10" | 5760001    | 5920000    |
| 7   | "chr10" | 5920001    | 6000000    |
| 8   | "chr10" | 6000001    | 7560000    |
| 9   | "chr10" | 7560001    | 9360000    |
| 10  | "chr10" | 9360001    | 11520000   |
⋮
```

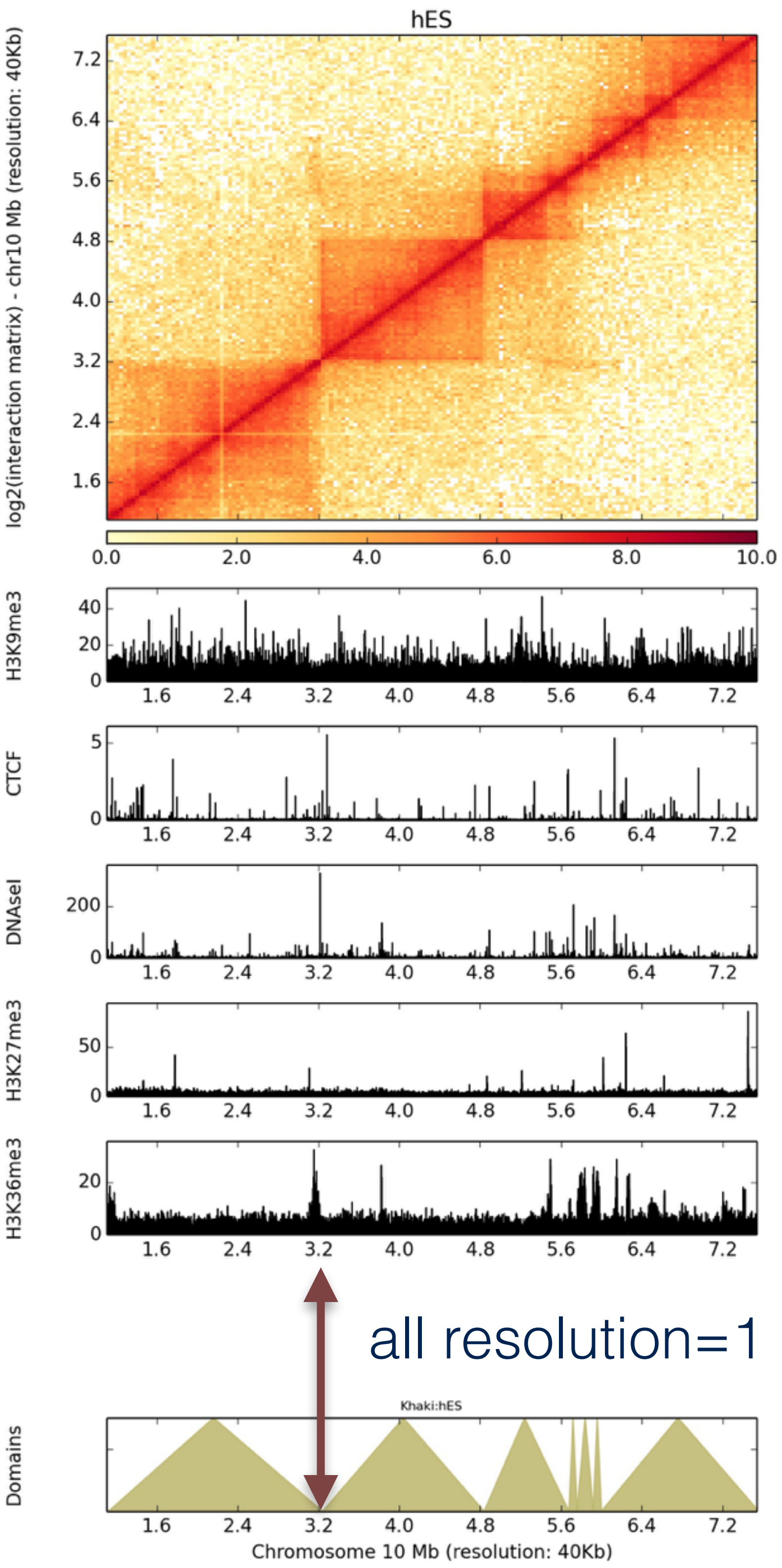




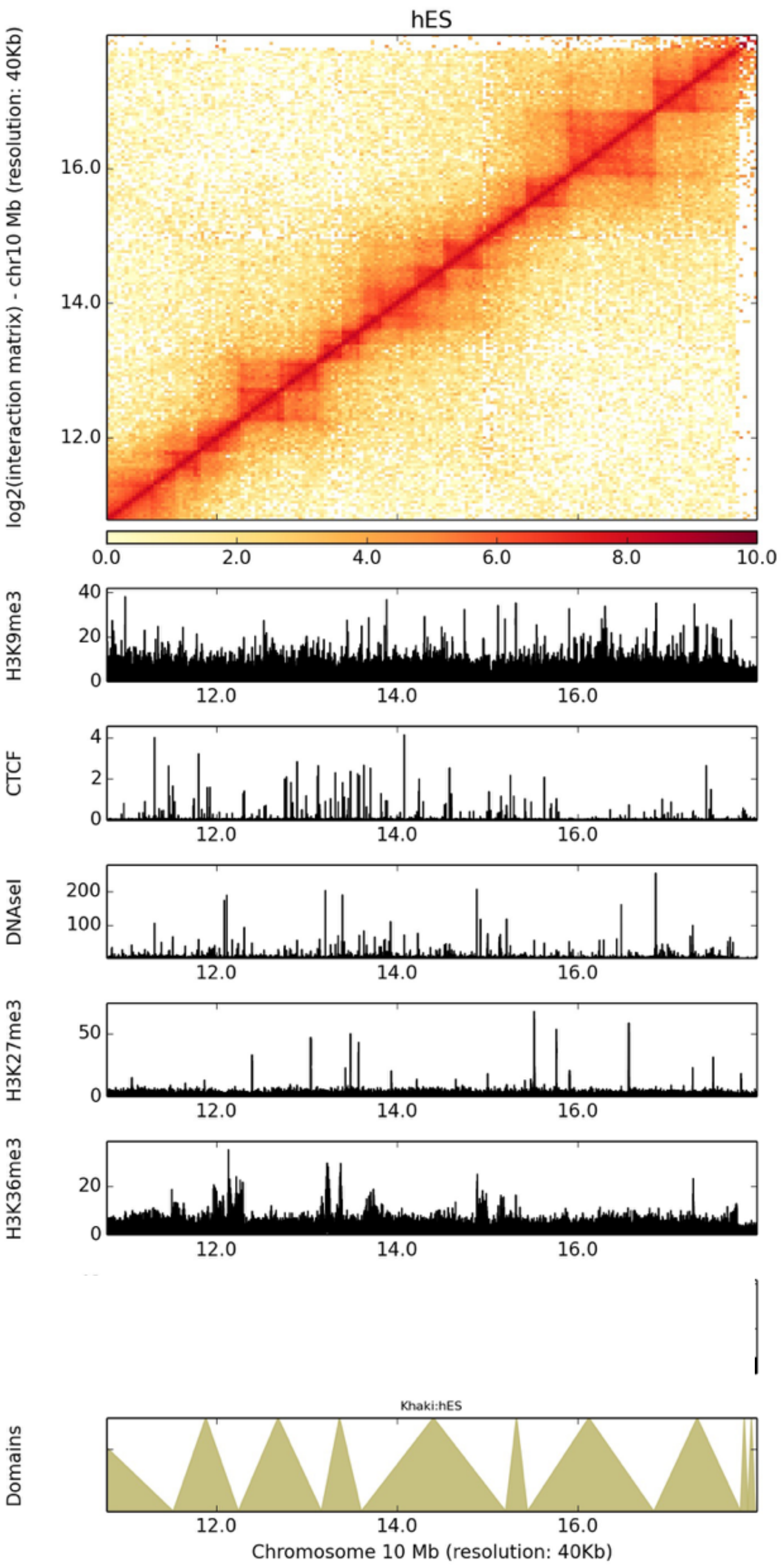

A
hES
log2(interaction matrix) - chr10 Mb (resolution: 40Kb)
36.0
34.0
32.0
30.0
28.0
26.0
24.0
22.0
0.0
2.0
4.0
6.0
8.0
10.0
Domains
res=1
res=.8
res=1.25
22.0
24.0
26.0
28.0
30.0
32.0
34.0
36.0
Chromosome 10 Mb (resolution: 40Kb)
B
number of TADs
6000
5000
4000
3000
2000
1000
0
0.6
0.8
1.0
1.2
1.4
1.6
1.8
resolution γ
C
size of TADs (log10)
8
7.5
7
6.5
6
5.5
5
0.625
0.75
0.875
1
1.125
1.25
1.375
1.5
1.625
1.75
resolution γ
D
normalized mutual information
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.6
0.8
1.0
1.2
1.4
1.6
1.8
resolution γ

chromatin features near domain boundary

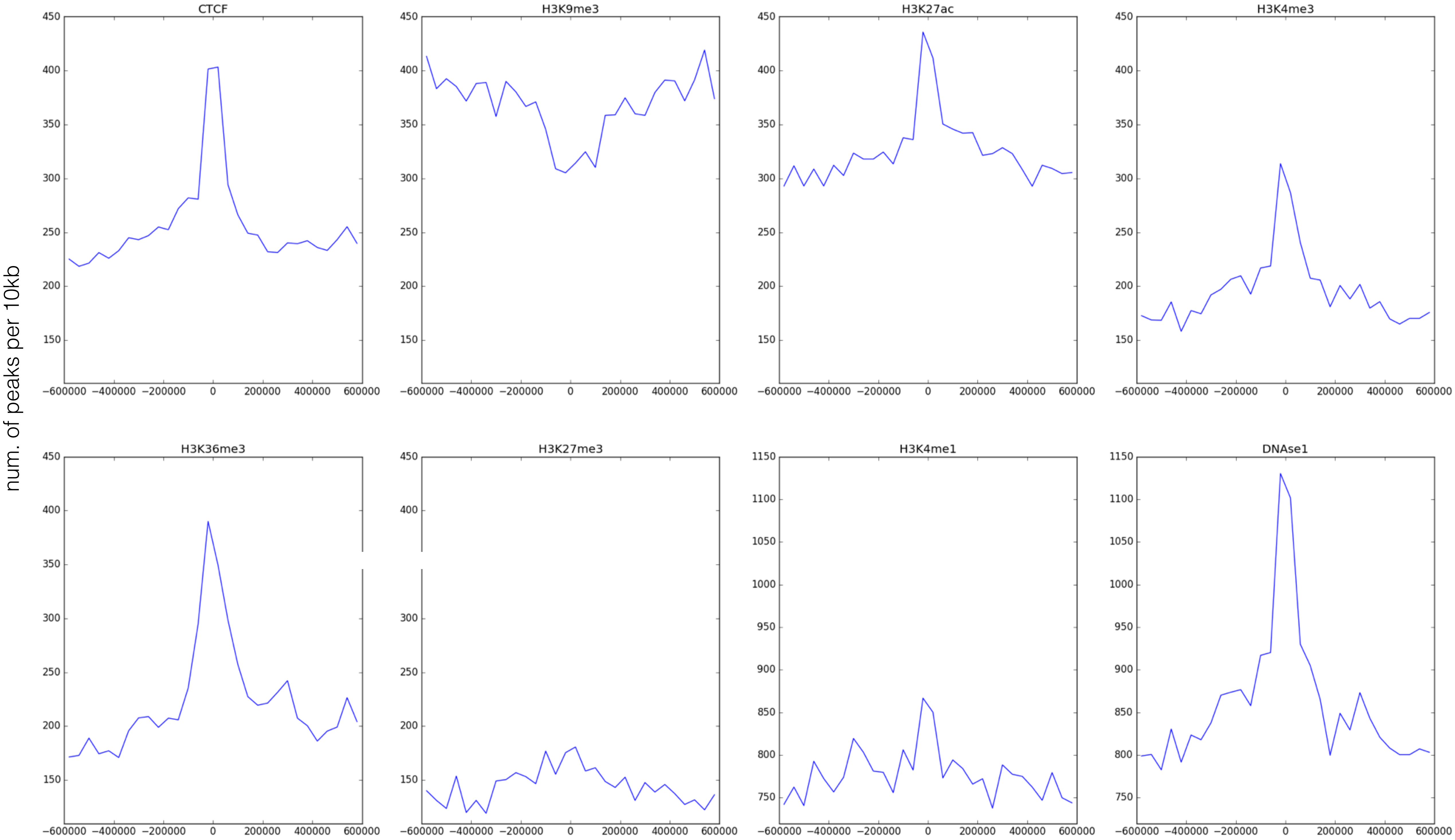




all resolution=1

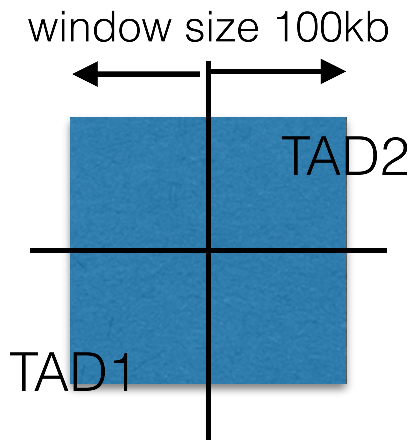


chromatin features near domain boundary



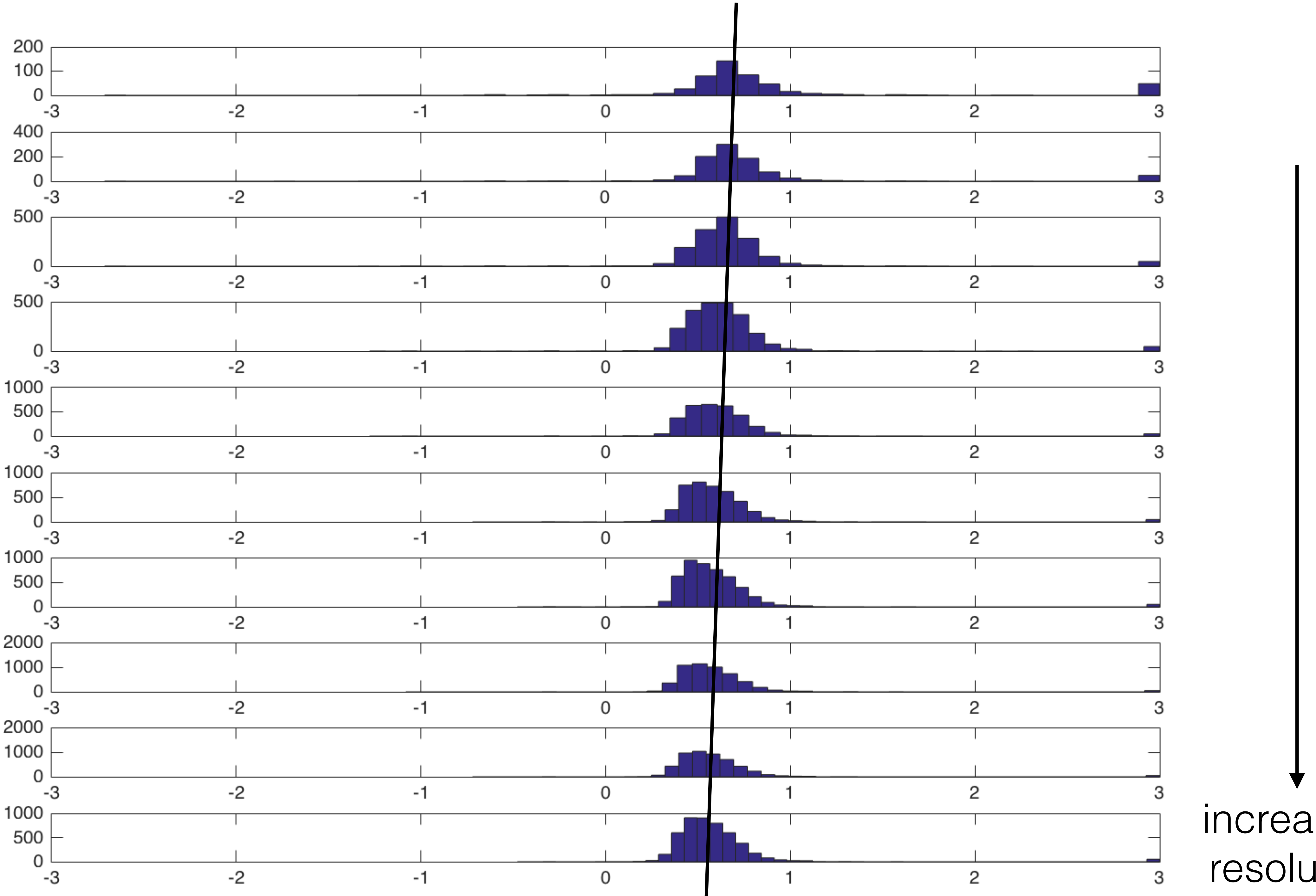
res=1, hESC, all chr

Domains interaction strength



interaction strength between 1 and 2 =

$$\frac{S11 + S22}{S12 + S21}$$

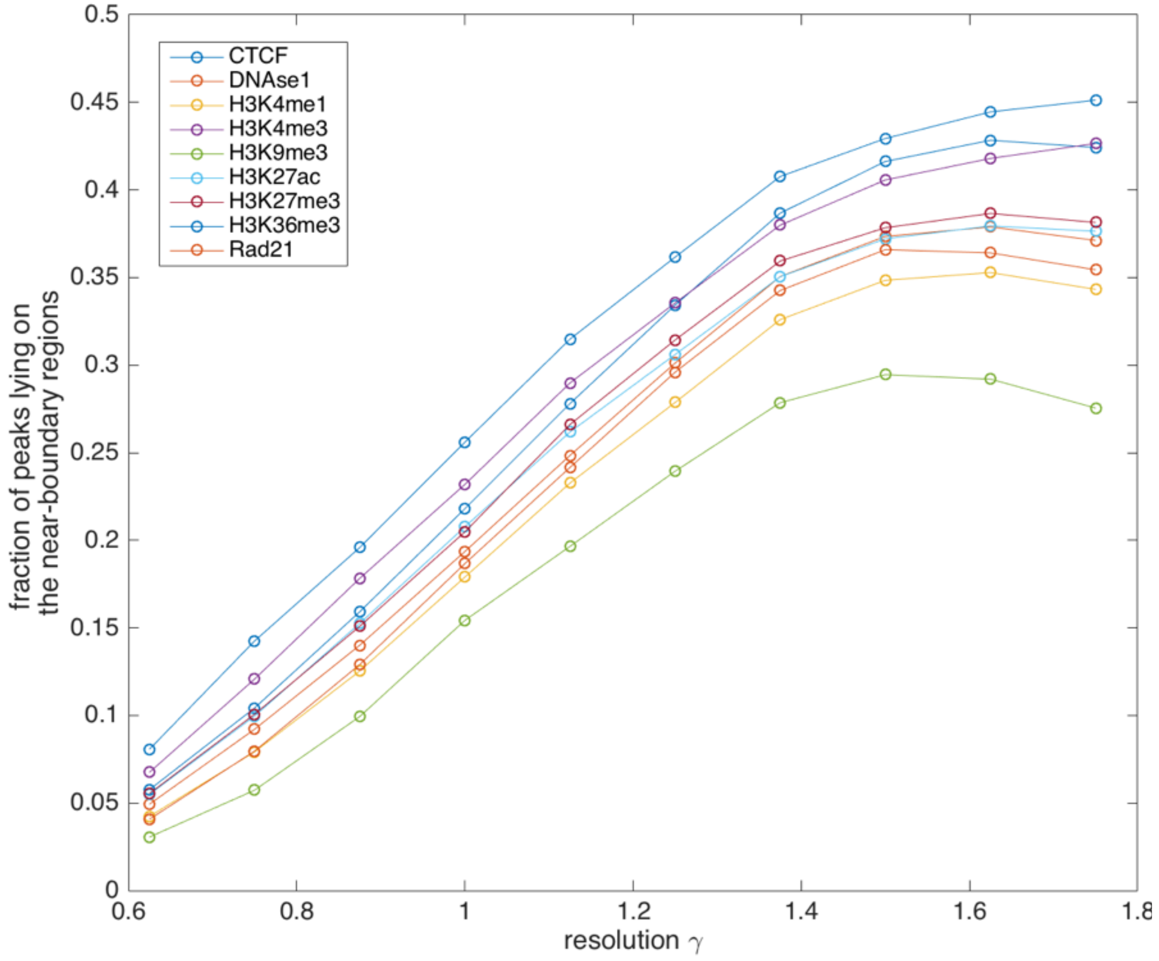


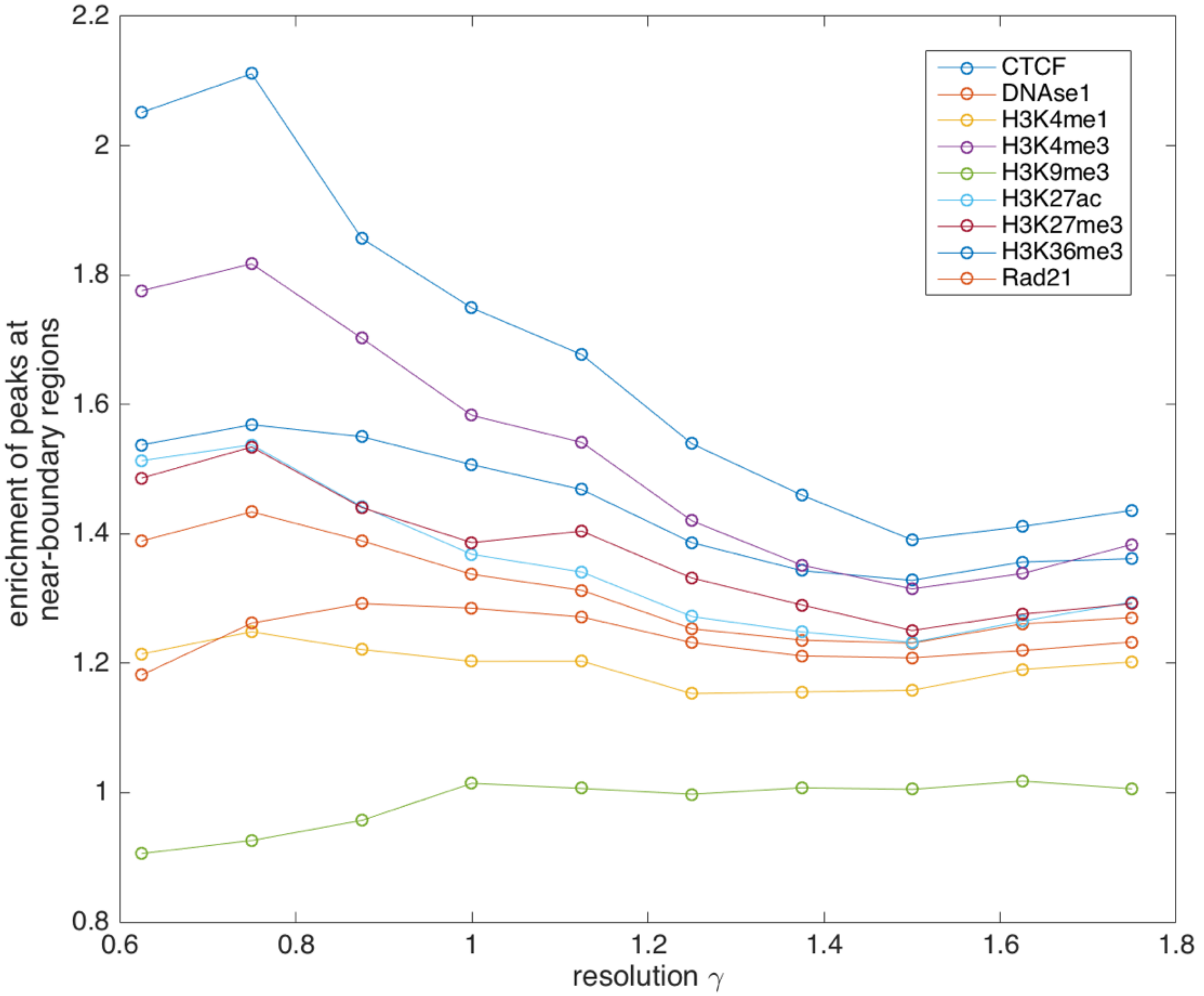
increasing resolution

interaction strength at boundaries (log10)

chromatin features near domain boundary

look at all peaks, what fraction of them are close to the TAD boundaries?





Identifying boundary regions based on histone mark

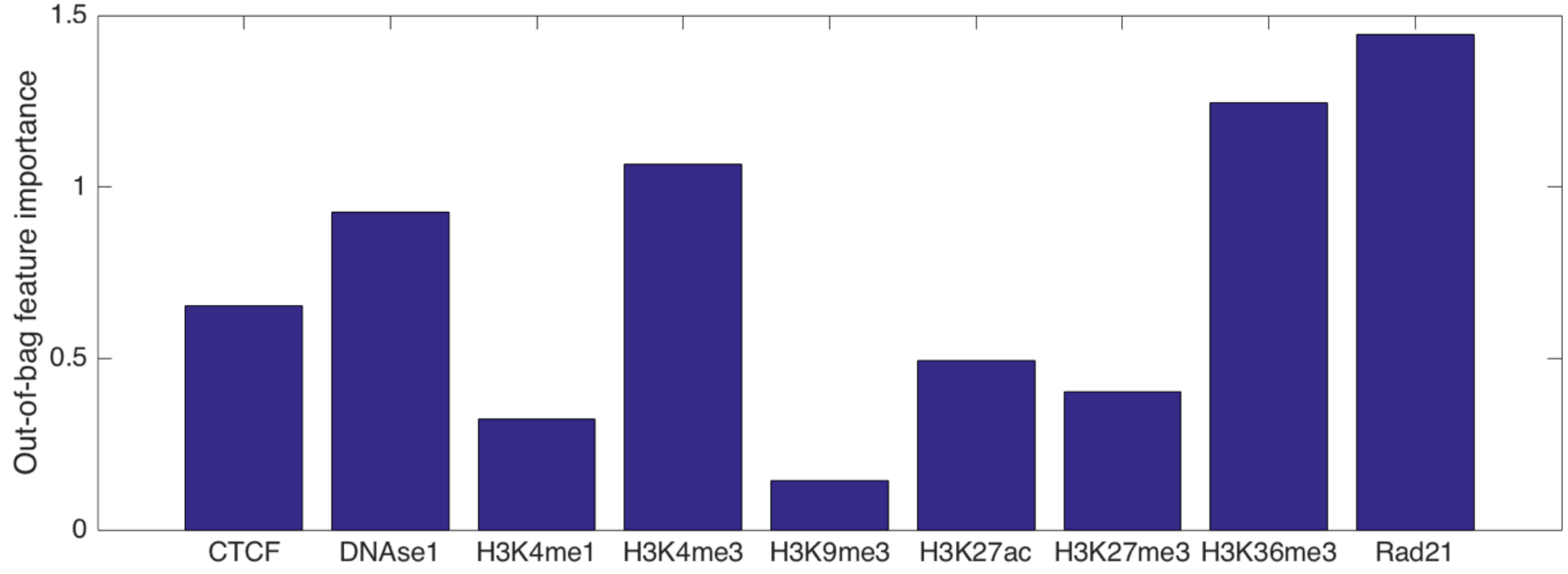
Classification

TAD boundary regions (40kb)

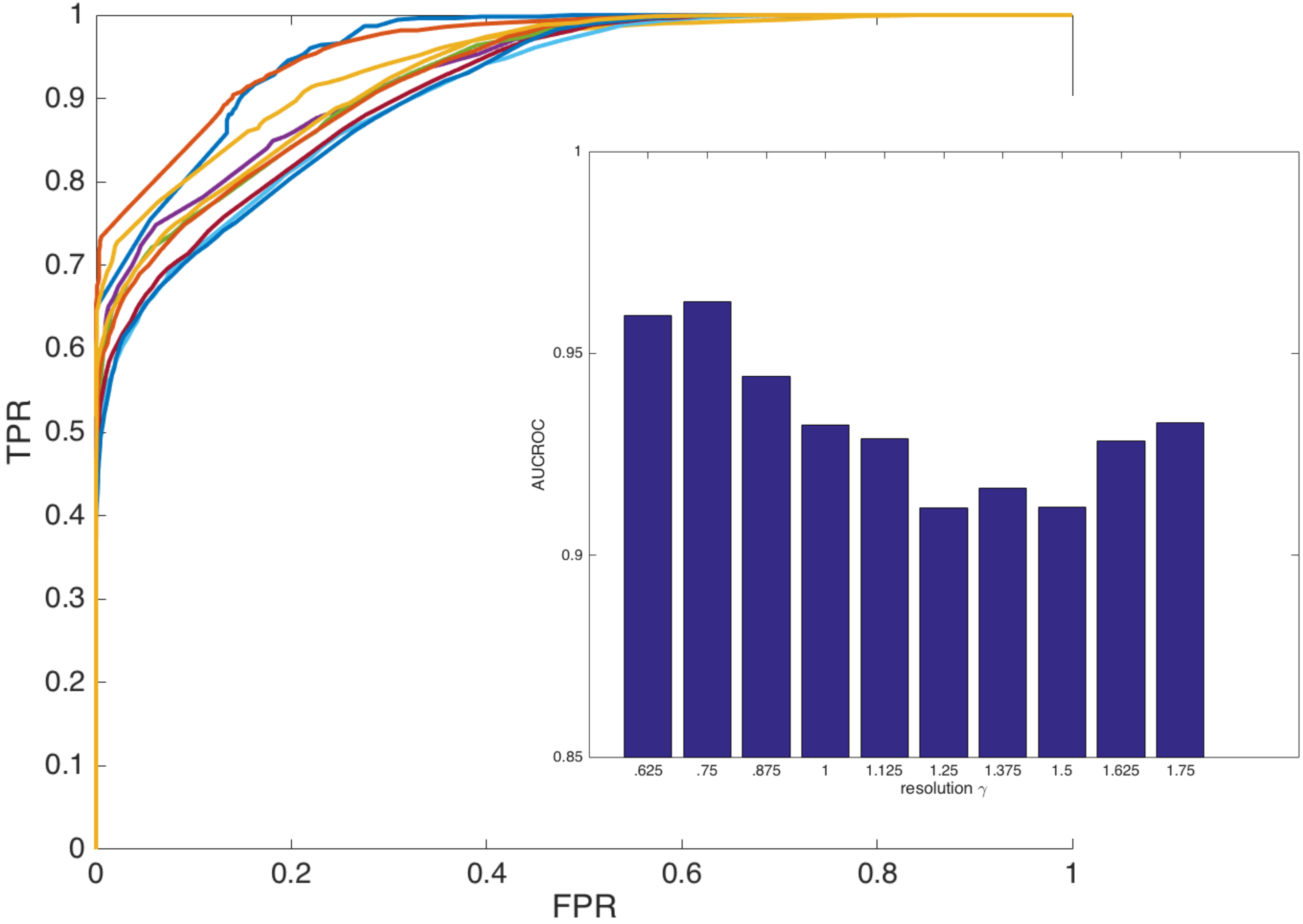
TAD middle regions (40kb)

Features
CTCF
DNAse1
H3K4me1
H3K4me3
H3K9me3
H3K27ac
H3K27me3
H3K36me3
Rad21

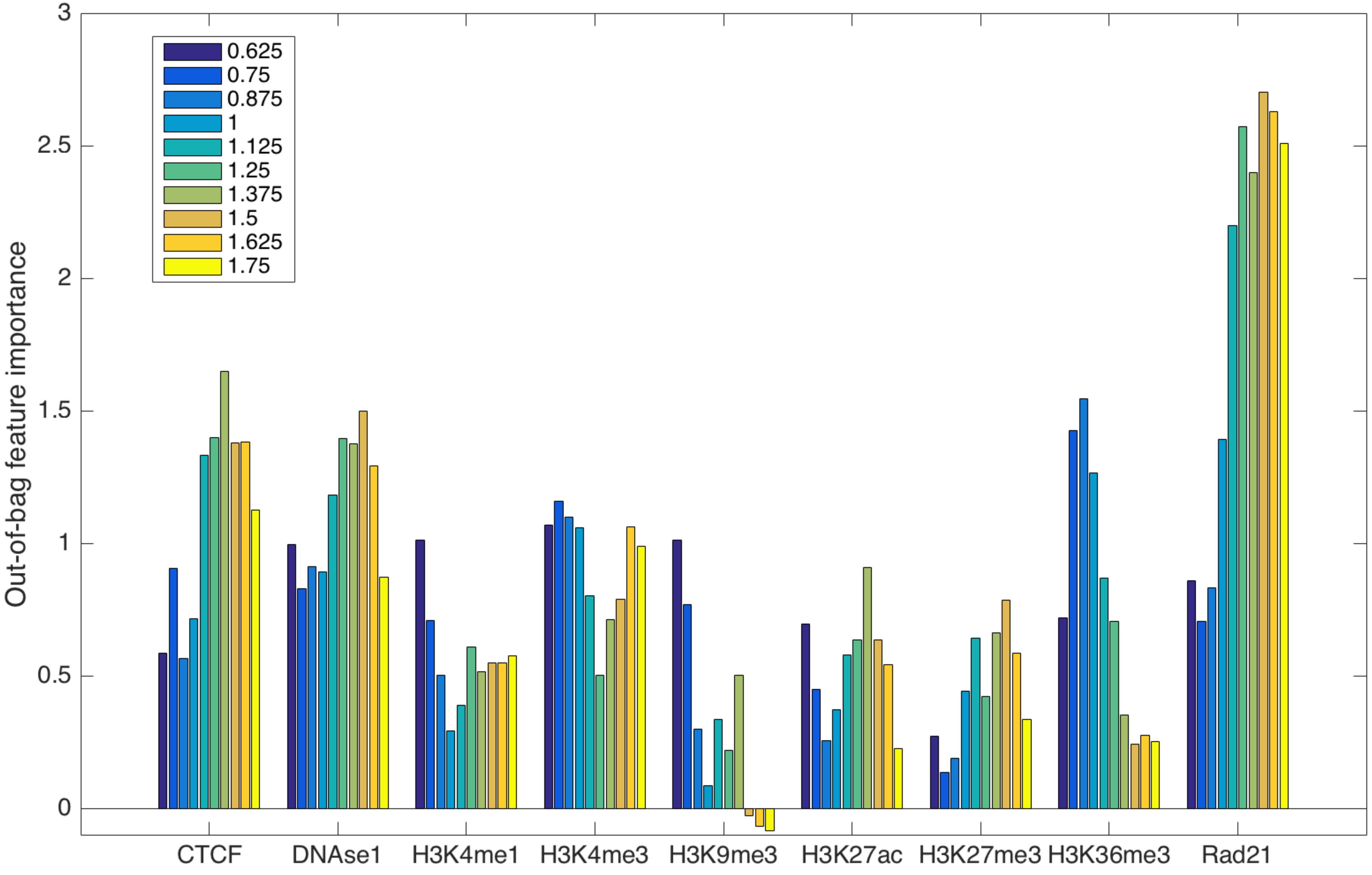
random forest AUC=0.93



Identifying boundary regions based on histone mark

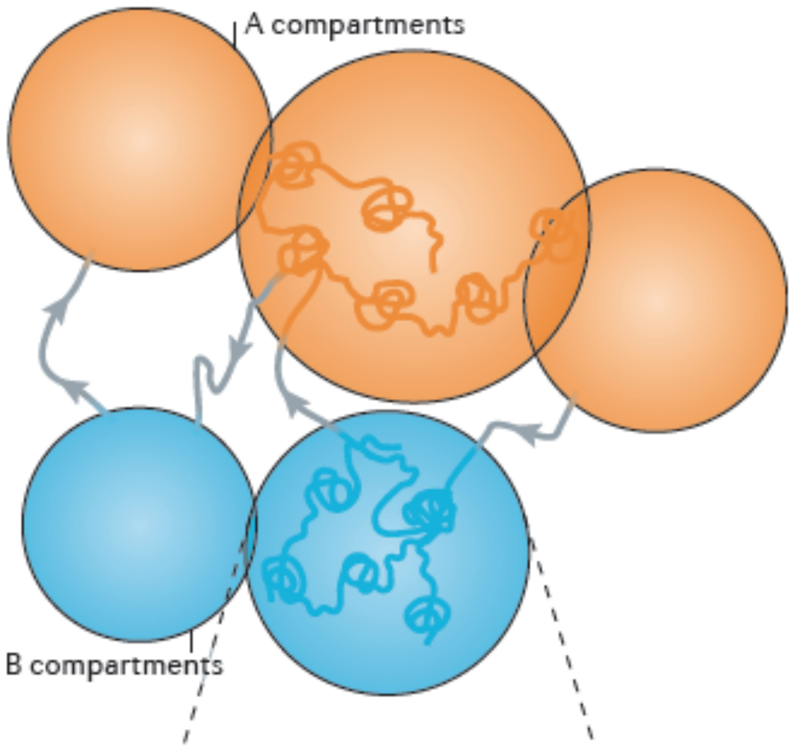


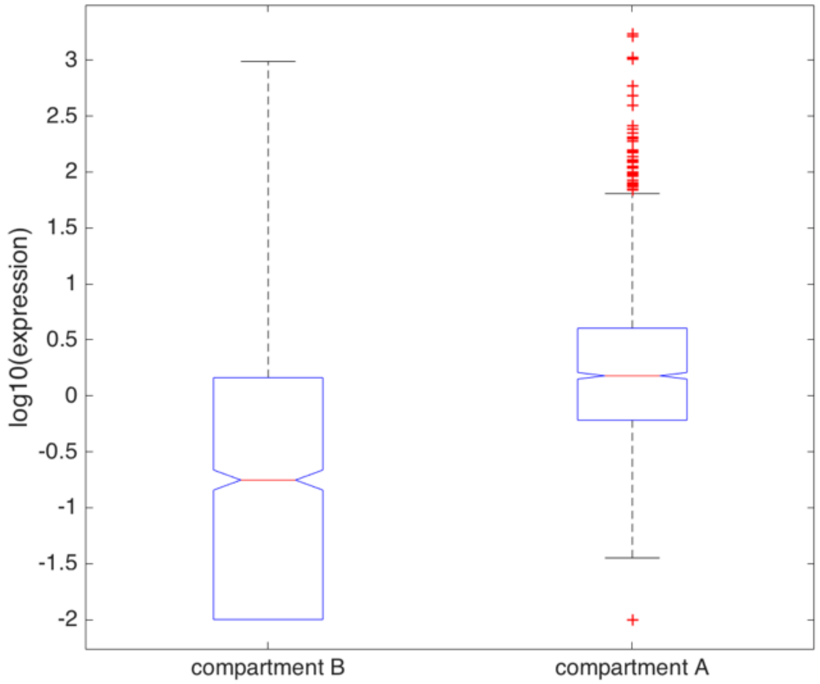
Identifying boundary regions based on histone mark

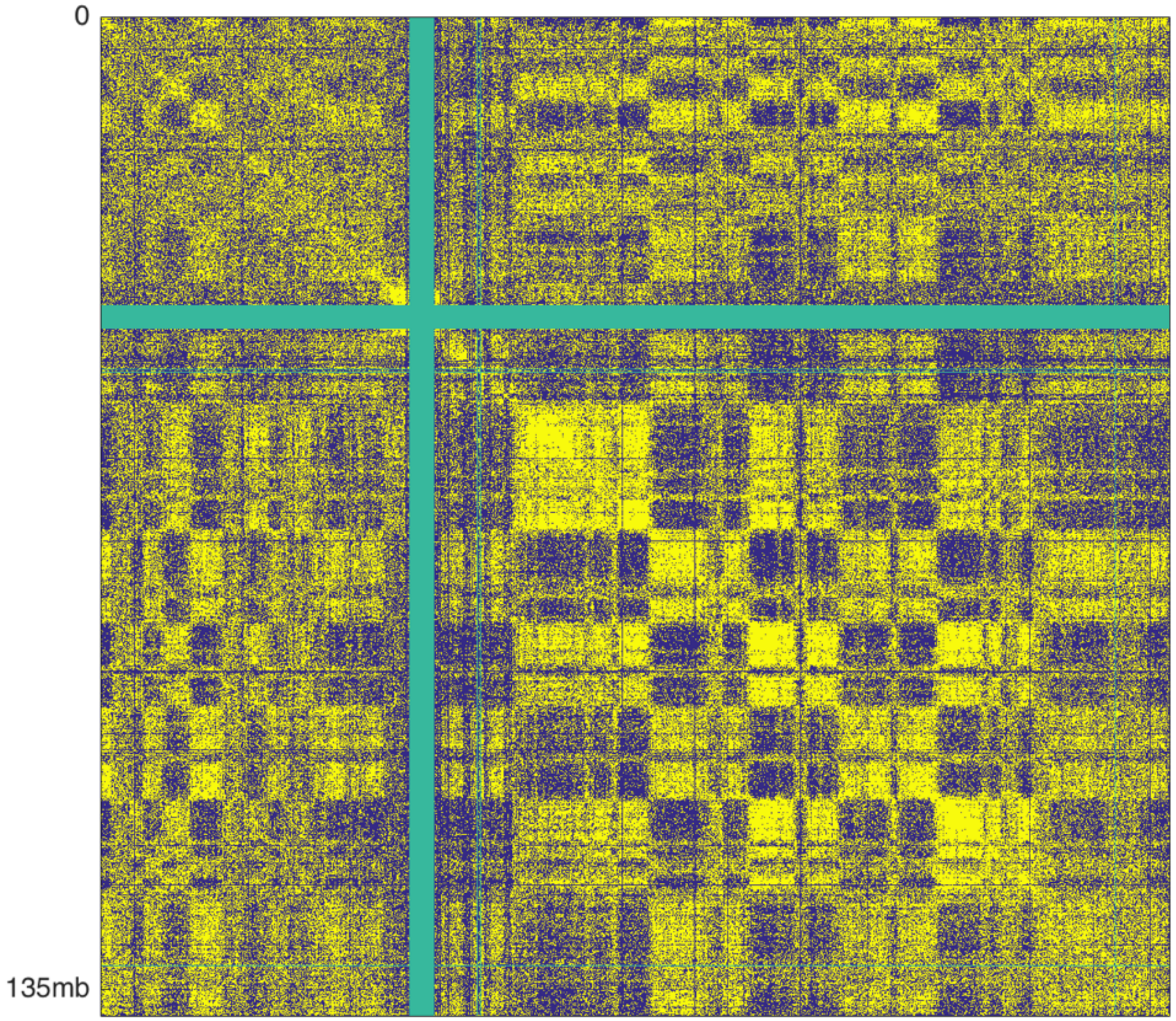


Compartments versus domains

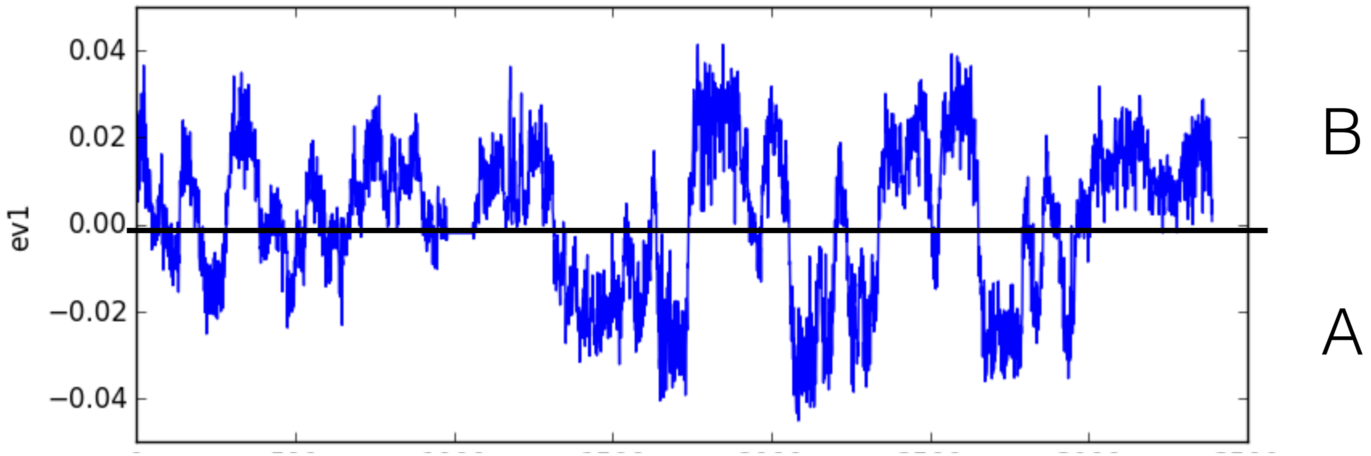
$$C_{ij} = cor(W_{ij}/E_{ij})$$



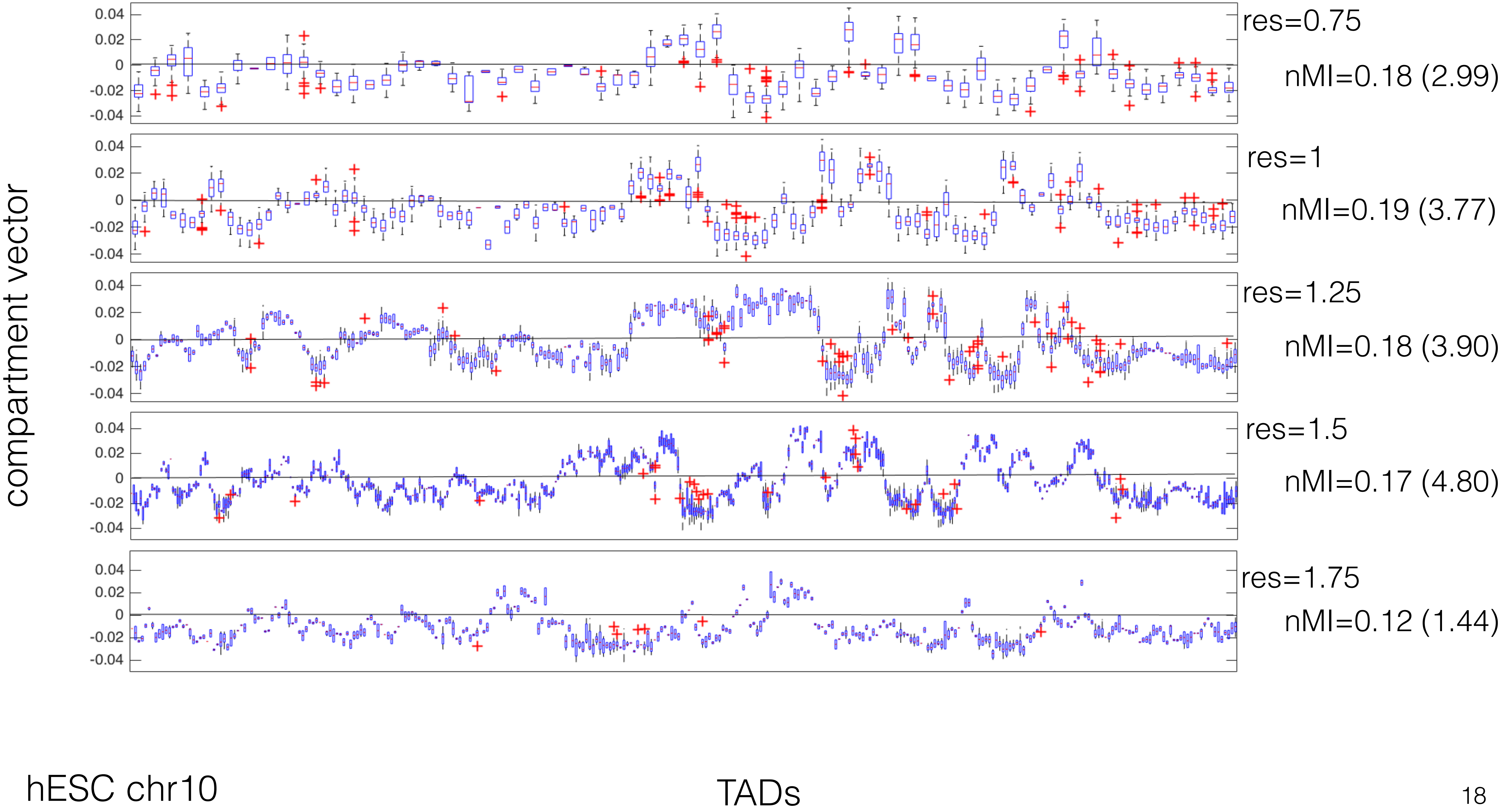




hESC chr10



Compartments versus domains



res=0.75

nMI=0.18 (2.99)

res=1

nMI=0.19 (3.77)

res=1.25

nMI=0.18 (3.90)

res=1.5

nMI=0.17 (4.80)

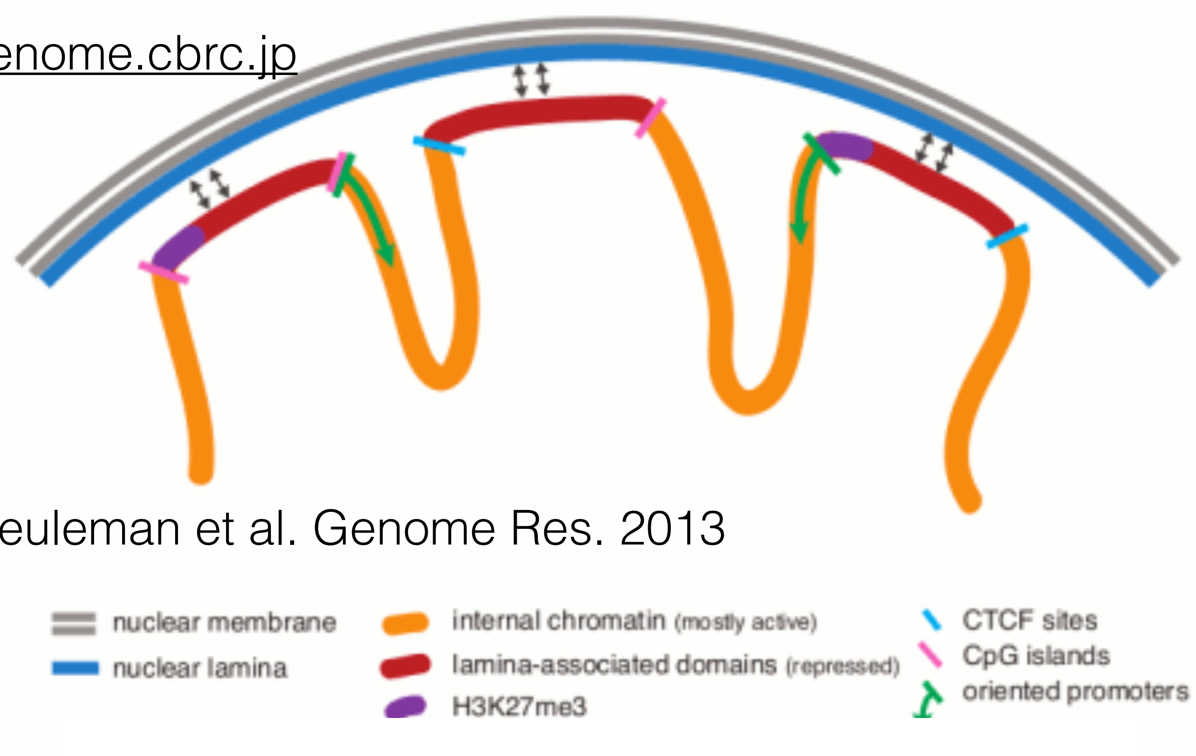
res=1.75

nMI=0.12 (1.44)

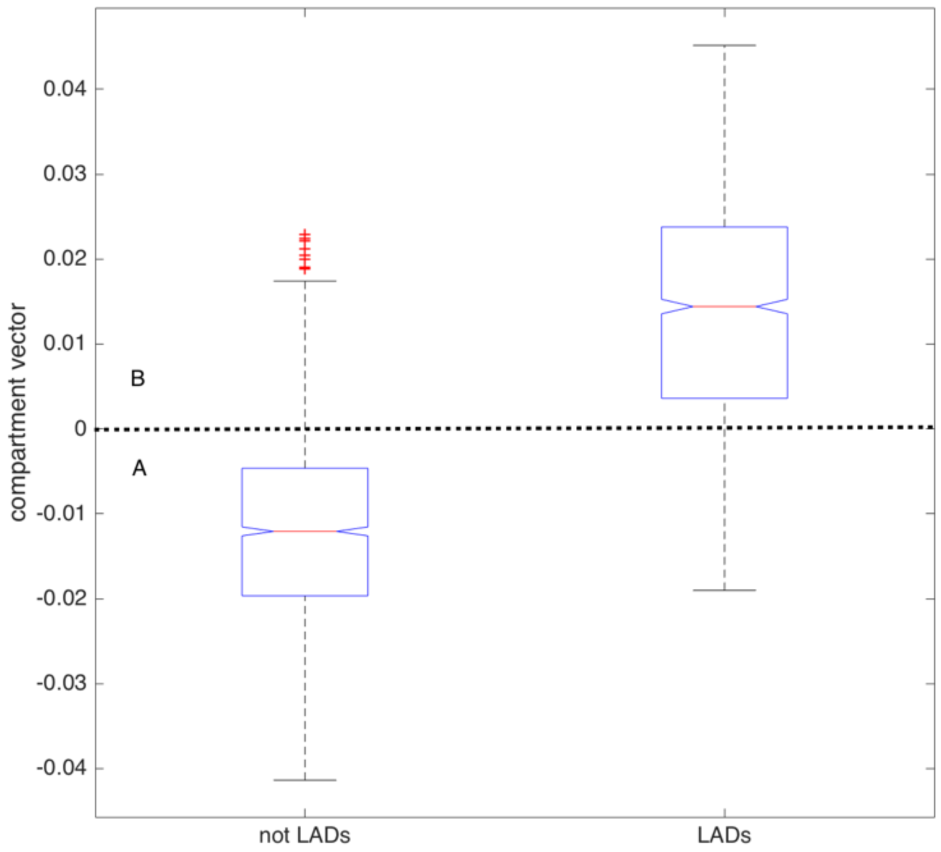
hESC chr10

Lamina associated domains (LADs)

epigenome.cbrc.jp



Meuleman et al. Genome Res. 2013



the next steps

- Keep fishing the interplay between TADs and other genome annotation in different resolution
  - annotation like ChromHMM
  - replication timing (ENCODE repli-seq)
  - TF binding pattern in TADs across multiple resolutions (with ANS, Yunsi)
  - k-mer frequency (ANS, Yunsi, SKL)
- Comparison between different cell types
- To compare with existing methods: Dixon et al. Nature 2012, Rao et al. Cell 2014, Weinreb and Raphael Bioinformatics 2015 (TADtree), Malik and Patro bioRxiv 2015 (Matryoshka)