messing around with Hi-C data

Koon-Kiu Yan

Group Meeting, March 2015

3D organization of genome

"We finished the genome map, now we can't figure out how to fold it."

image credit: Iyer et al. BMC Biophysics 2011, cartoonist John Chase

3D organization of genome

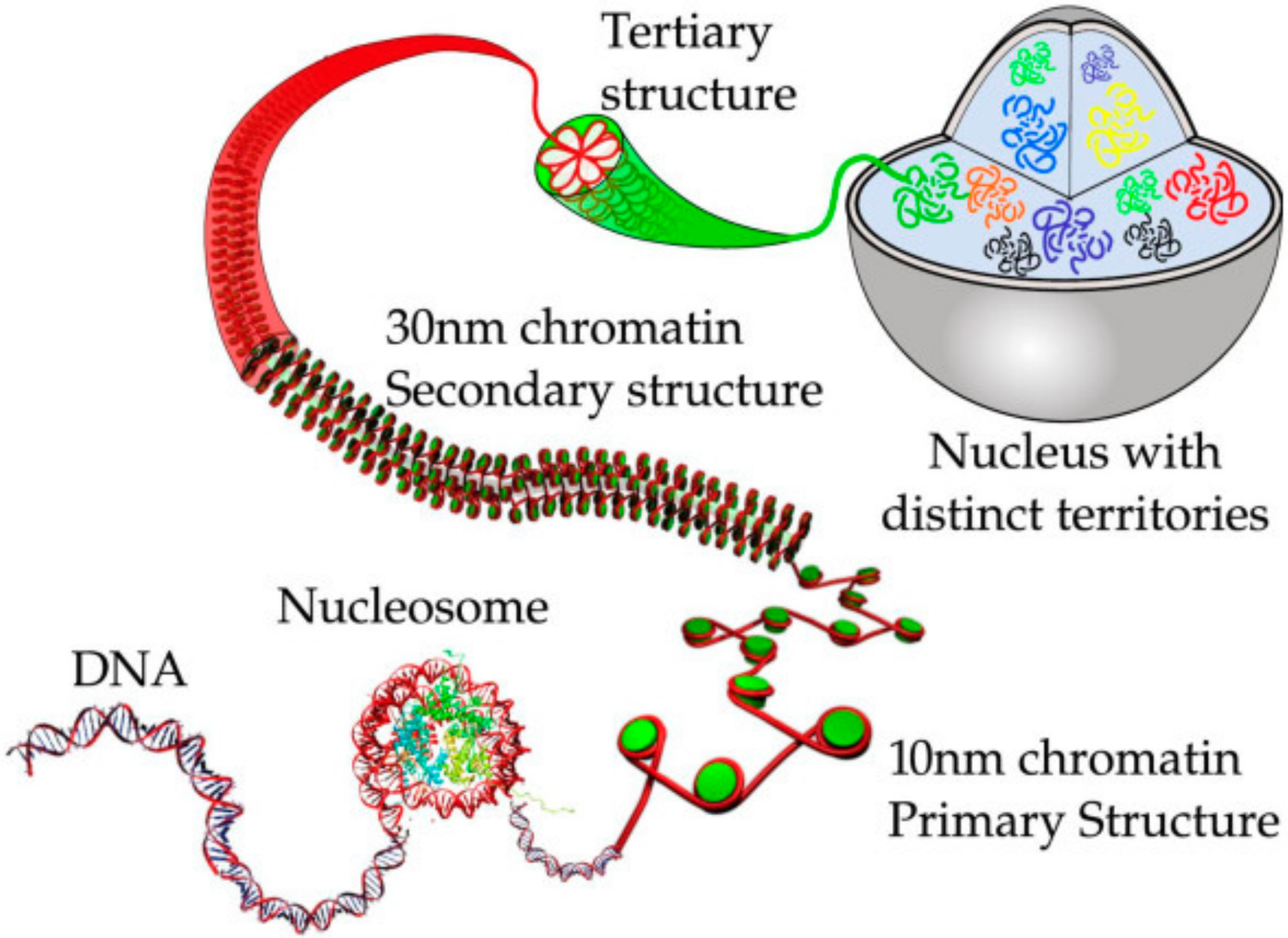


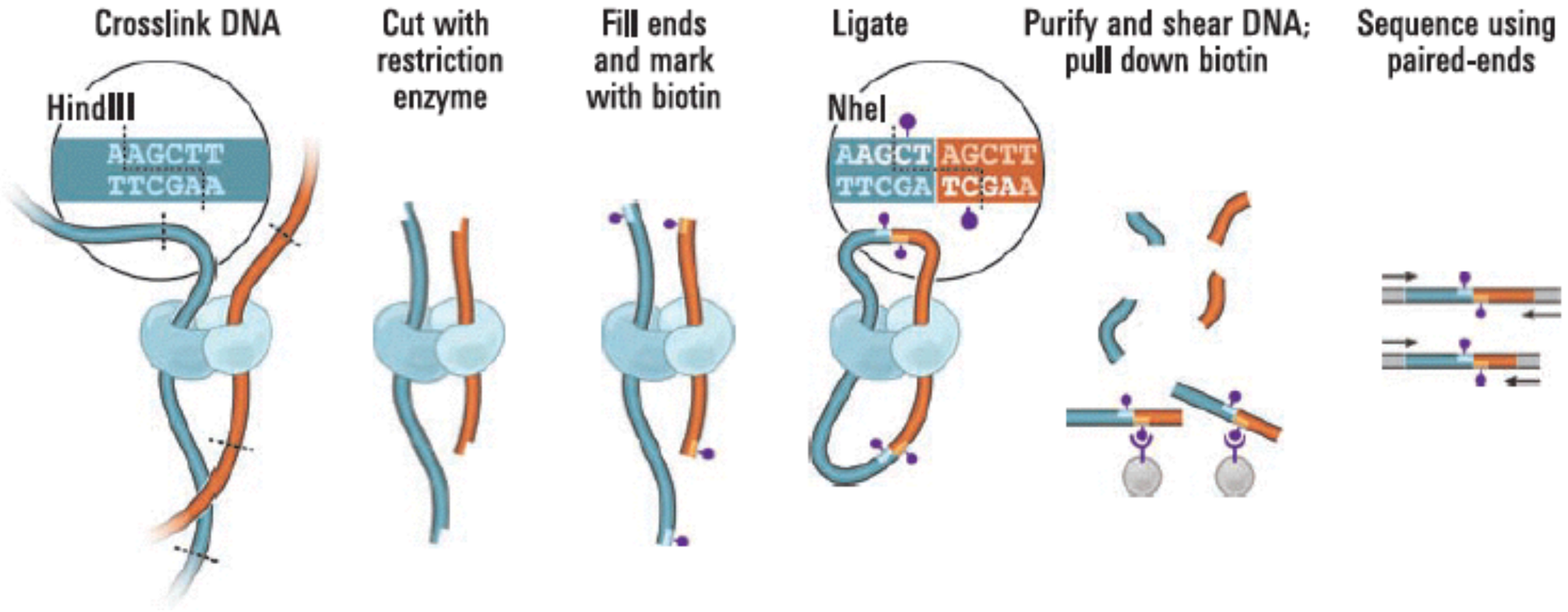
image credit: Iyer et al. BMC Biophysics 2011

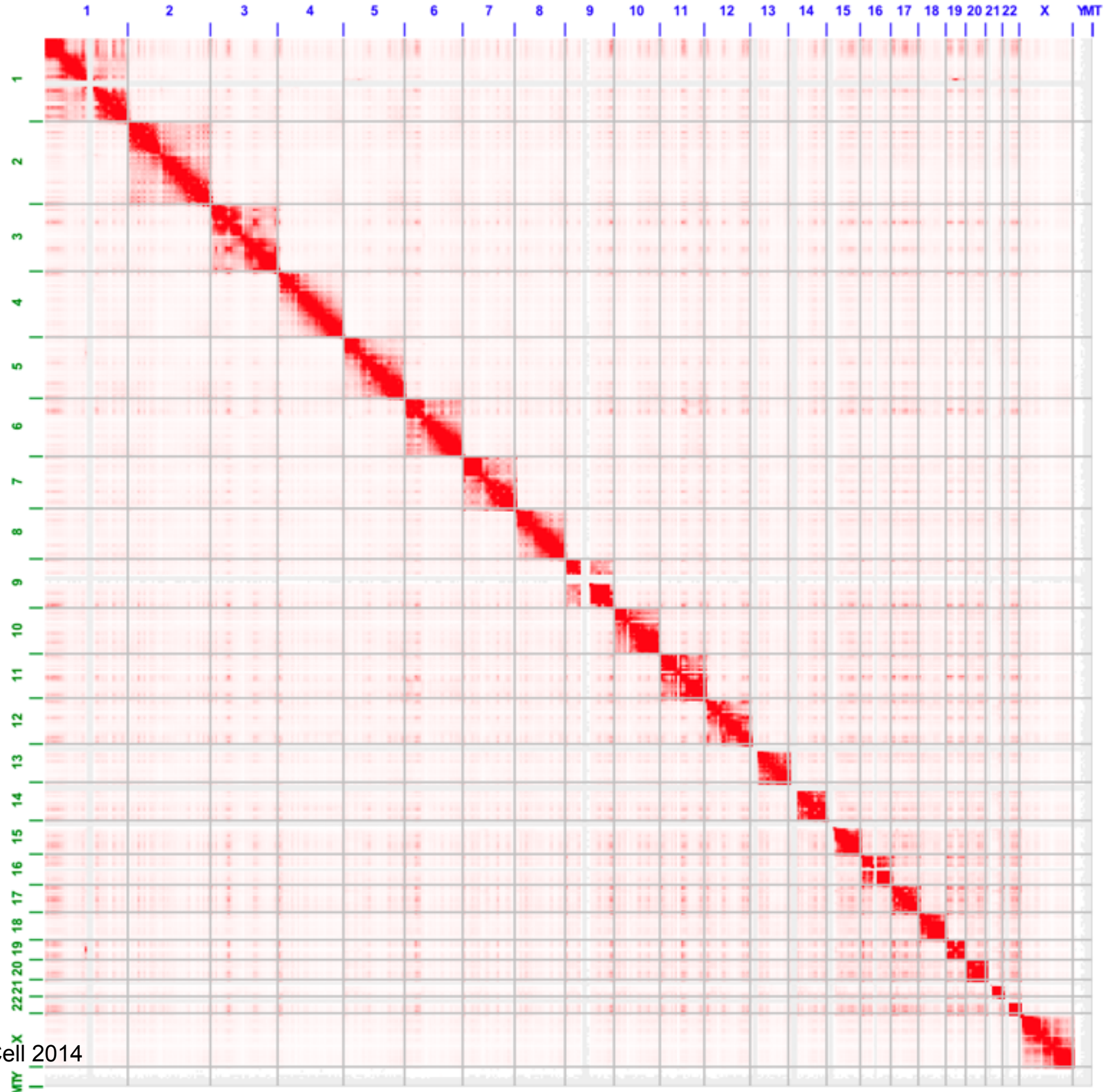
Chromosome conformation capture (3C) and Hi-C

Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome

Erez Lieberman-Aiden,[1,2,3,4*] Nynke L. van Berkum,[5*] Louise Williams,[1] Maxim Imakaev,[2] Tobias Ragoczy,[6,7] Agnes Telling,[6,7] Ido Amit,[1] Bryan R. Lajoie,[5] Peter J. Sabo,[8] Michael O. Dorschner,[8] Richard Sandstrom,[8] Bradley Bernstein,[1,9] M. A. Bender,[10] Mark Groudine,[6,7] Andreas Gnirke,[1] John Stamatoyannopoulos,[8] Leonid A. Mirny,[2,11] Eric S. Lander,[1,12,13]† Job Dekker[5]†

SCIENCE VOL 326 9 OCTOBER 2009





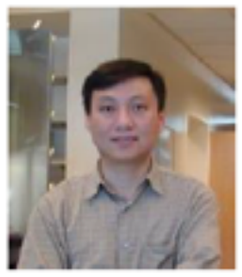
Data:
Rao et al. Aiden, Cell 2014

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

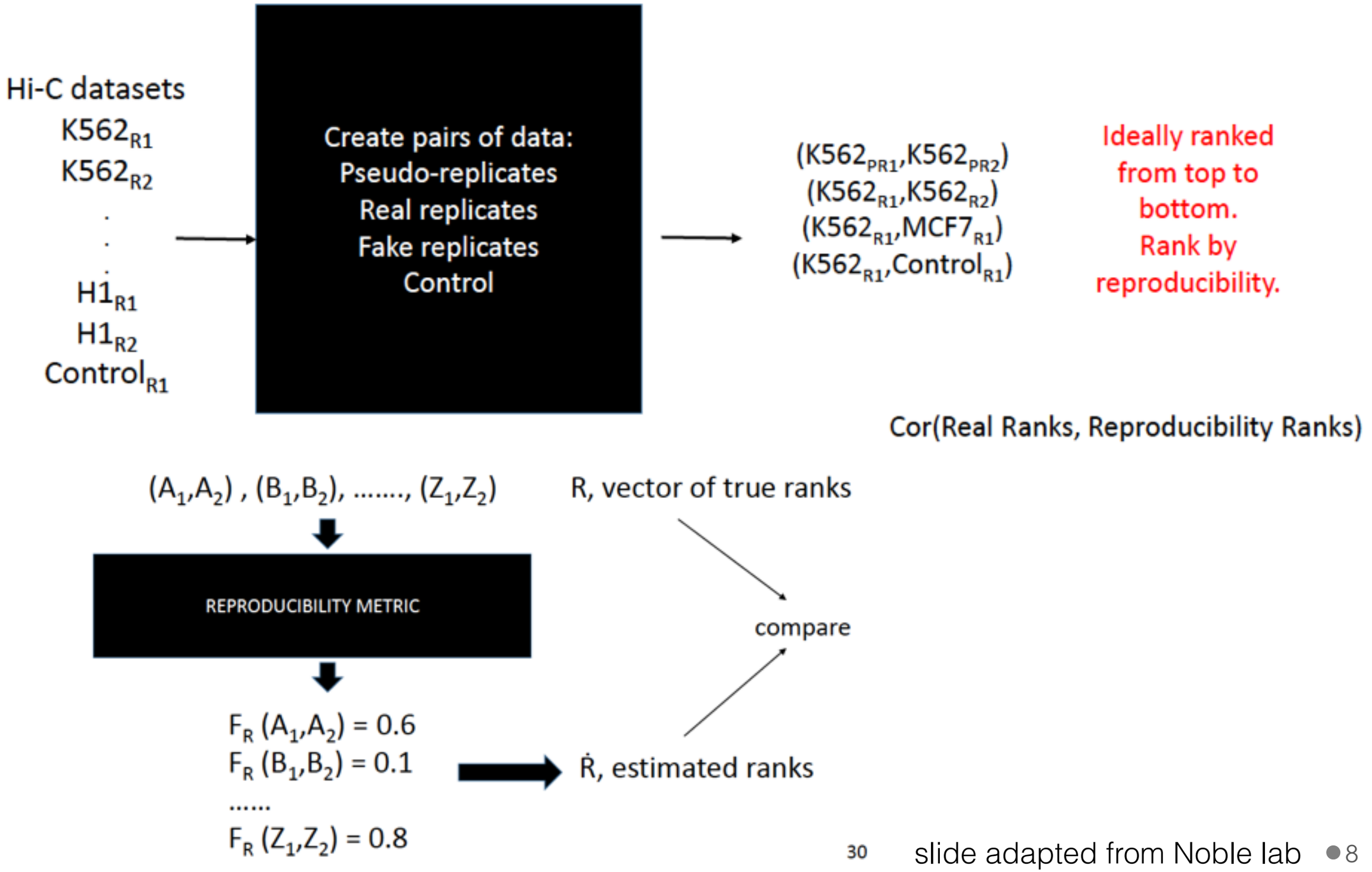
- Preparation of manuscript for ENCODE guidelines for assessing the quality and the reproducibility of chromosome conformation capture experiments
  - Similar to ENCODE ChIP-seq guidelines (Landt et al. Genome Research 2012)

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



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

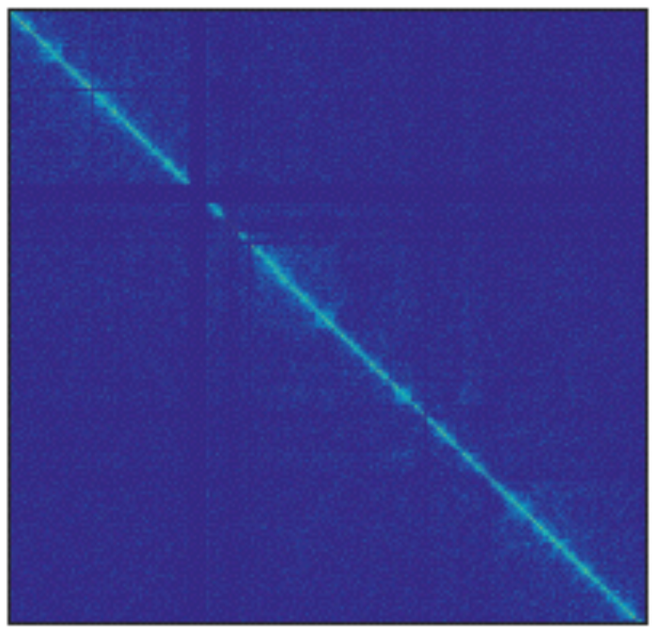
Evaluate reproducibility metric

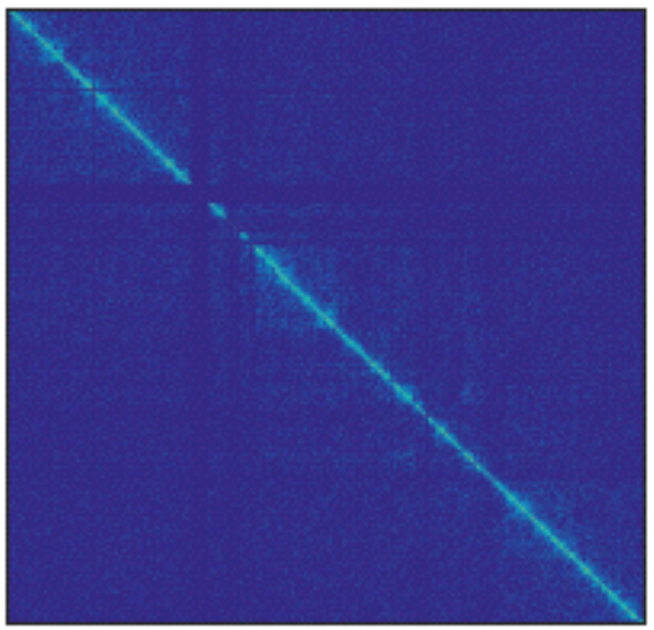


slide adapted from Noble lab

Comparing contact matrices is a technical challenge

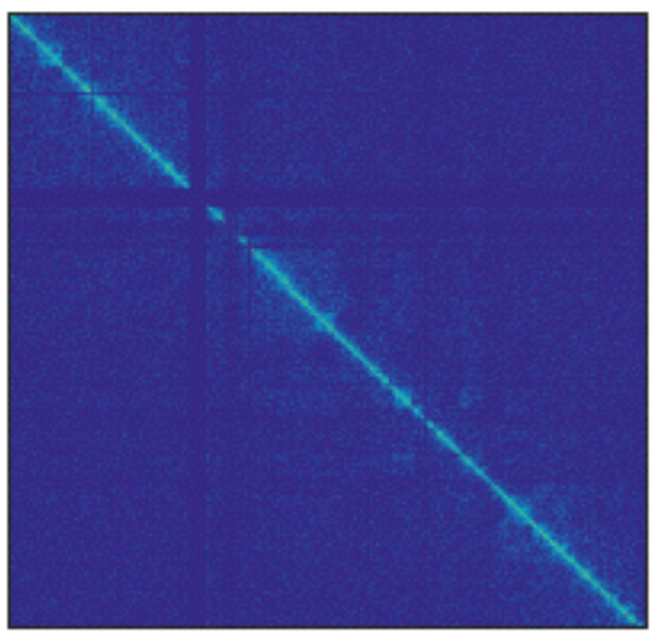
Pair 6

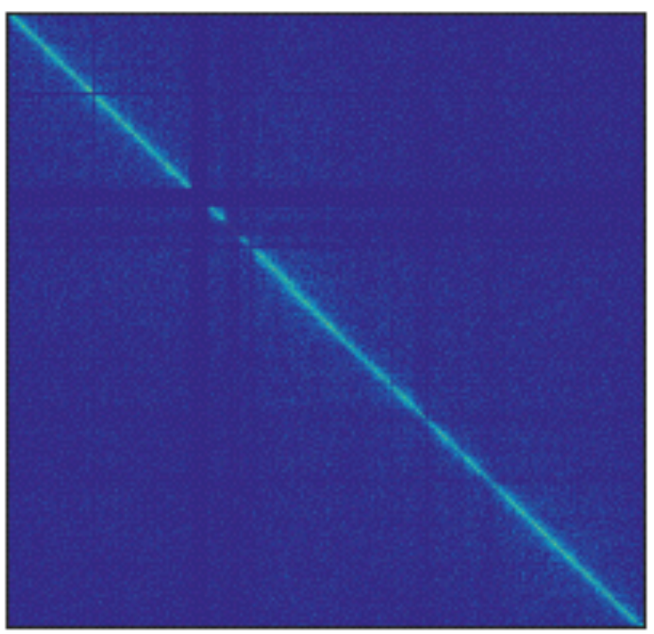




r=0.95

Pair 22





r=0.70

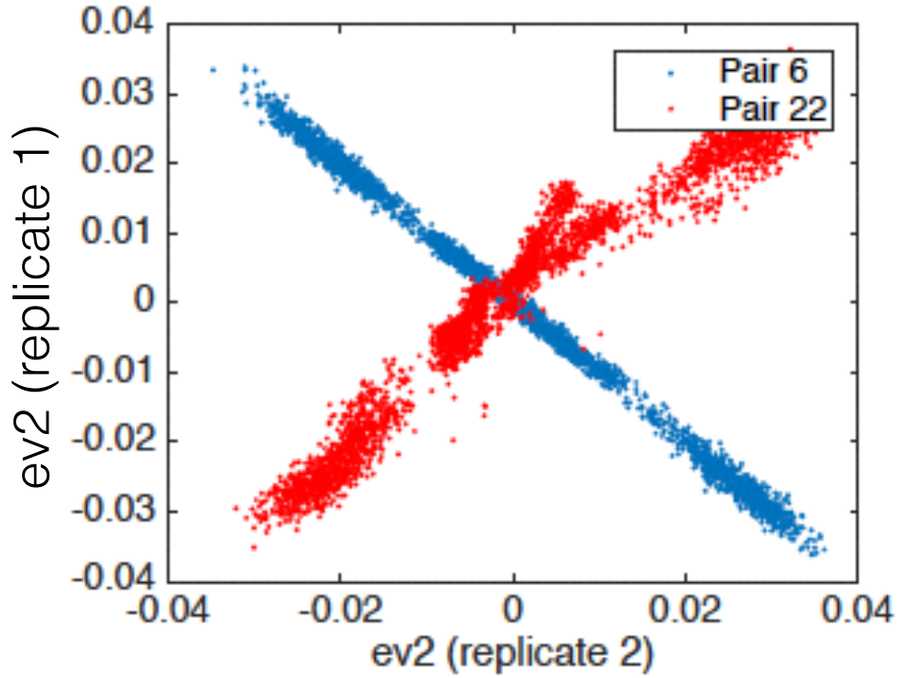
Quantifying reproducibility using spectral graph theory

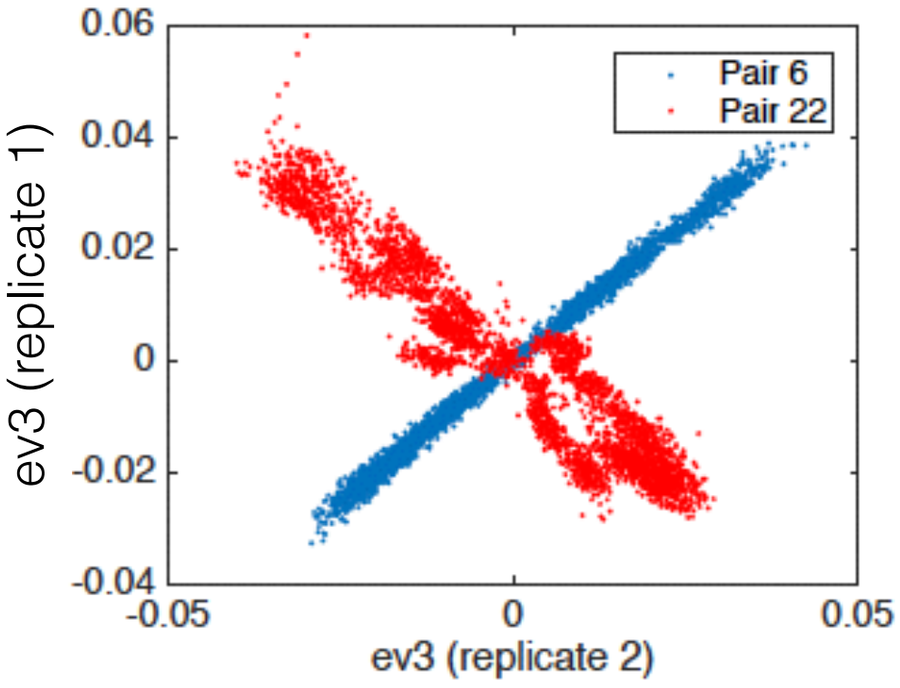
Laplacian $L = D - A$

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

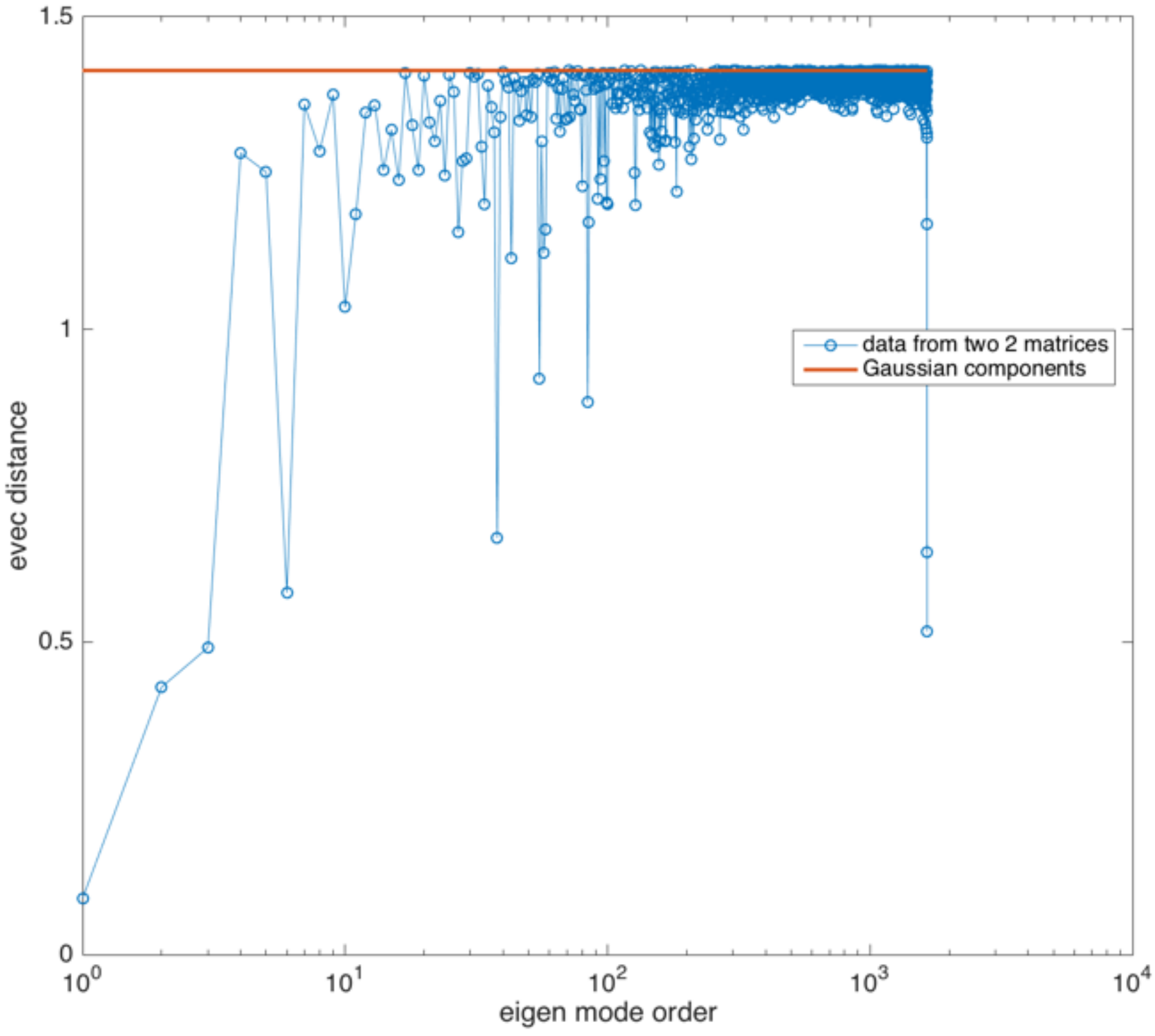
$$0 = \lambda_0 \leq \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$$

- leading eigenvectors capture the structures of the graph (dimension reduction)
- equivalent to eigenmodes of the corresponding random walk on the graph
- for each pair of leading eigenvectors, calculate the Euclidean distance

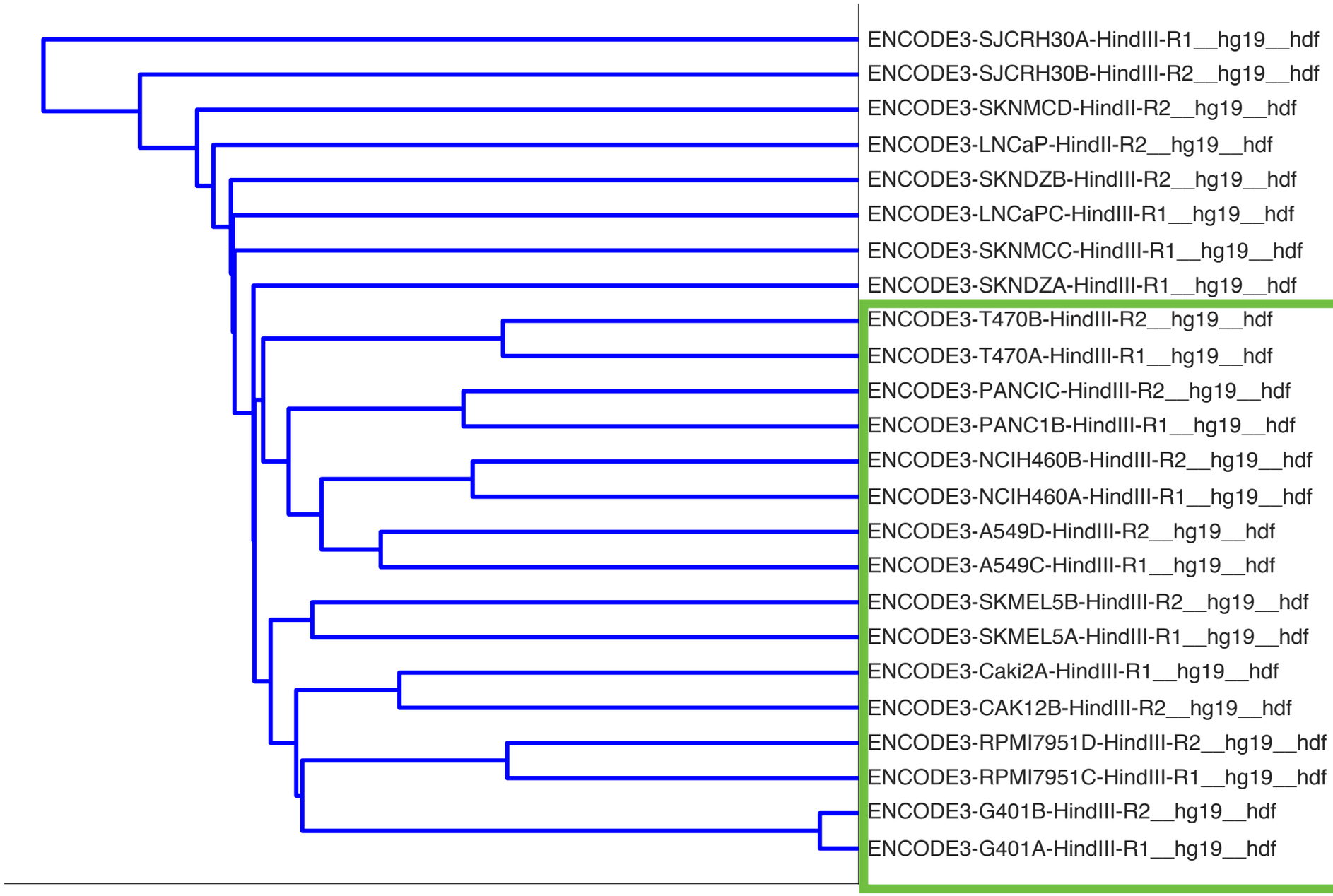




Quantifying reproducibility using spectral graph theory



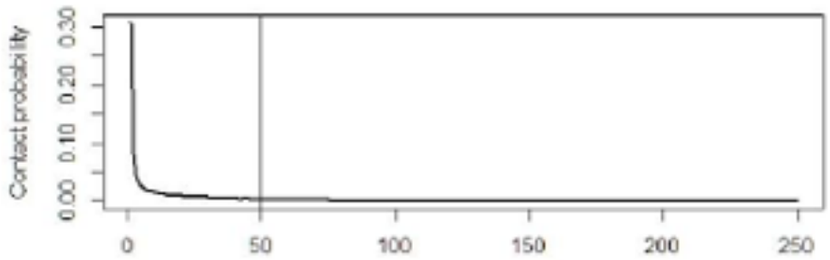
Quantifying reproducibility using spectral graph theory

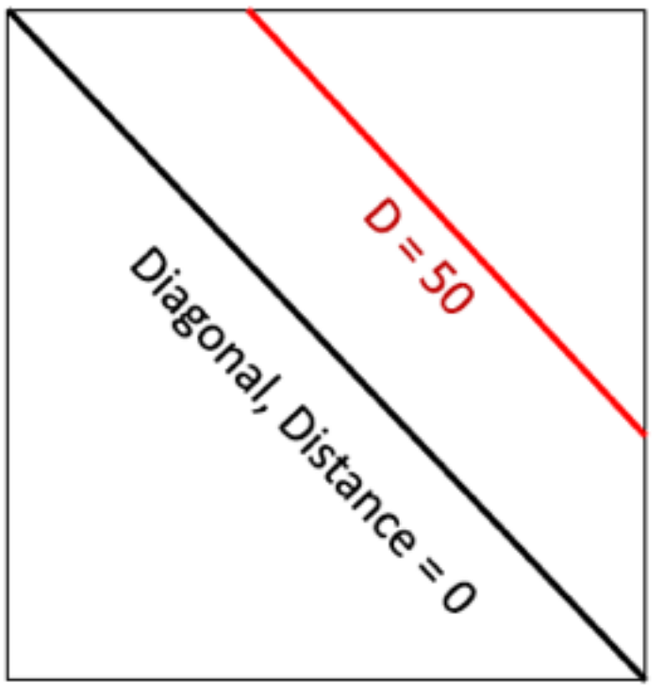


Simulating Noise in Hi-C

Distance effect

Step 1) Pick a distance 'D'

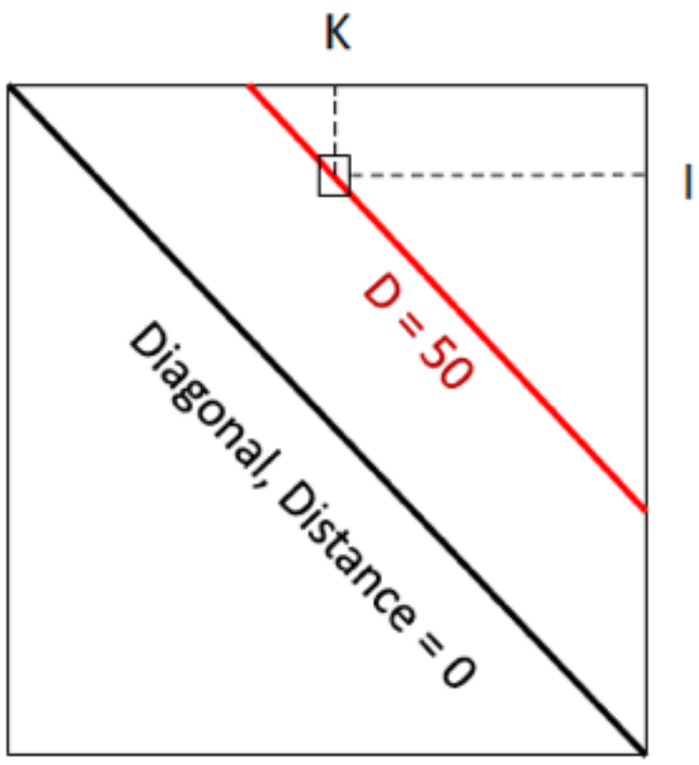




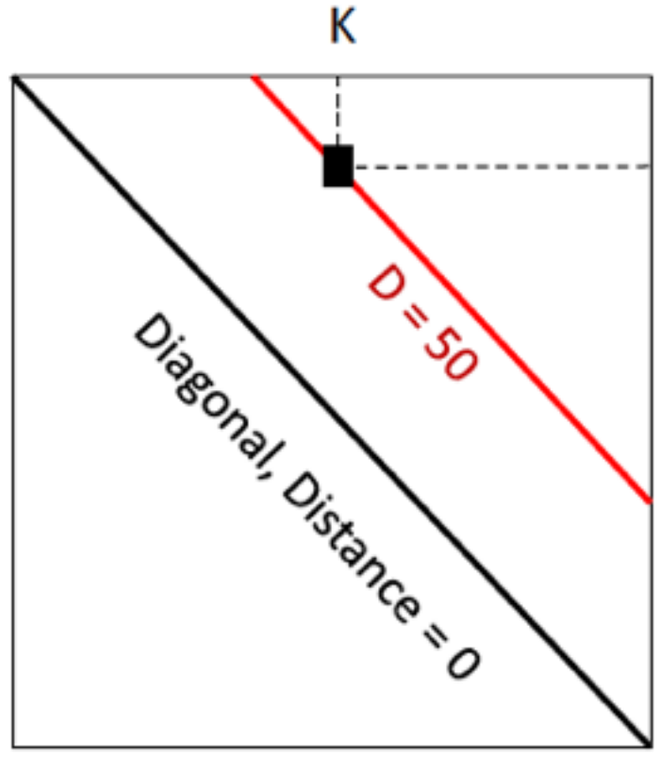
Mappability & GC biases

Step 2) Choose M_{IK} at distance 'D'

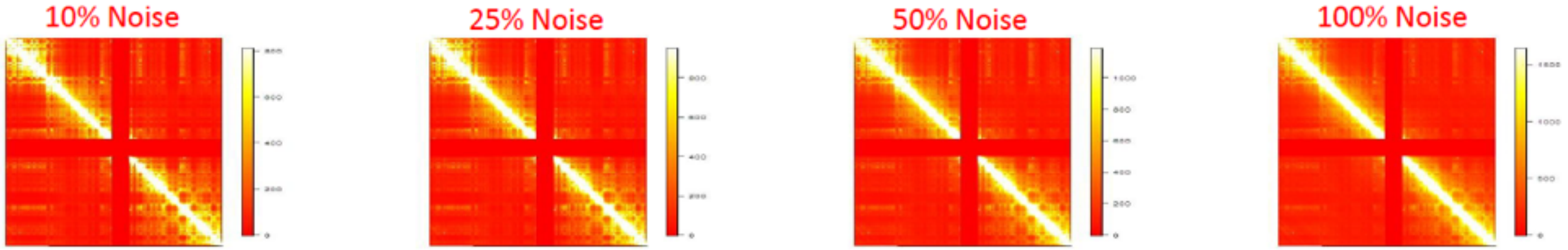
$P(I,K) \sim P(I) \times P(K)$



Step 3) Add +1 to chosen bin



Simulating Noise in Hi-C



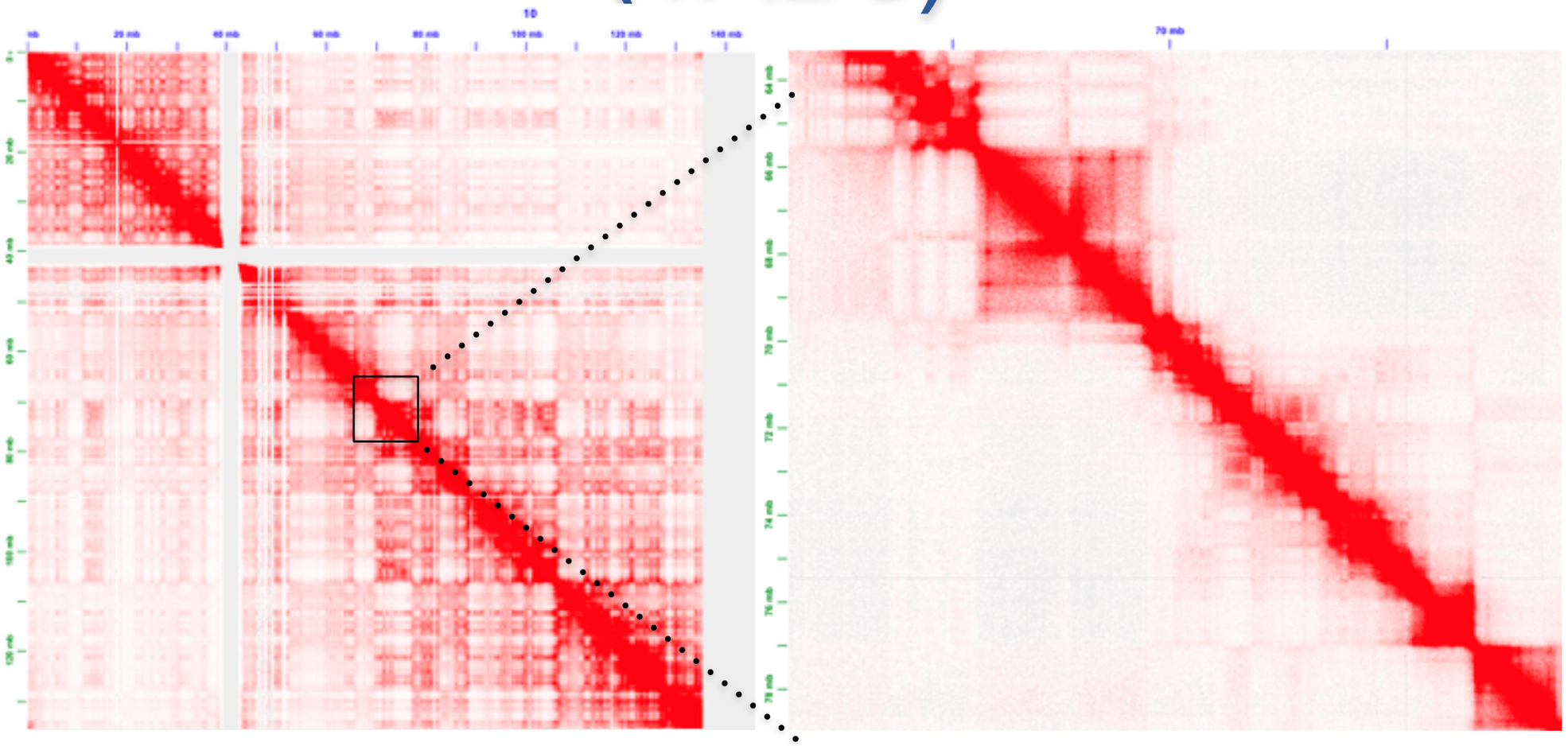
D, D+10%Noise, .. D+100%Noise

QUALITY CONTROL METRIC

slide adapted from Noble lab

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

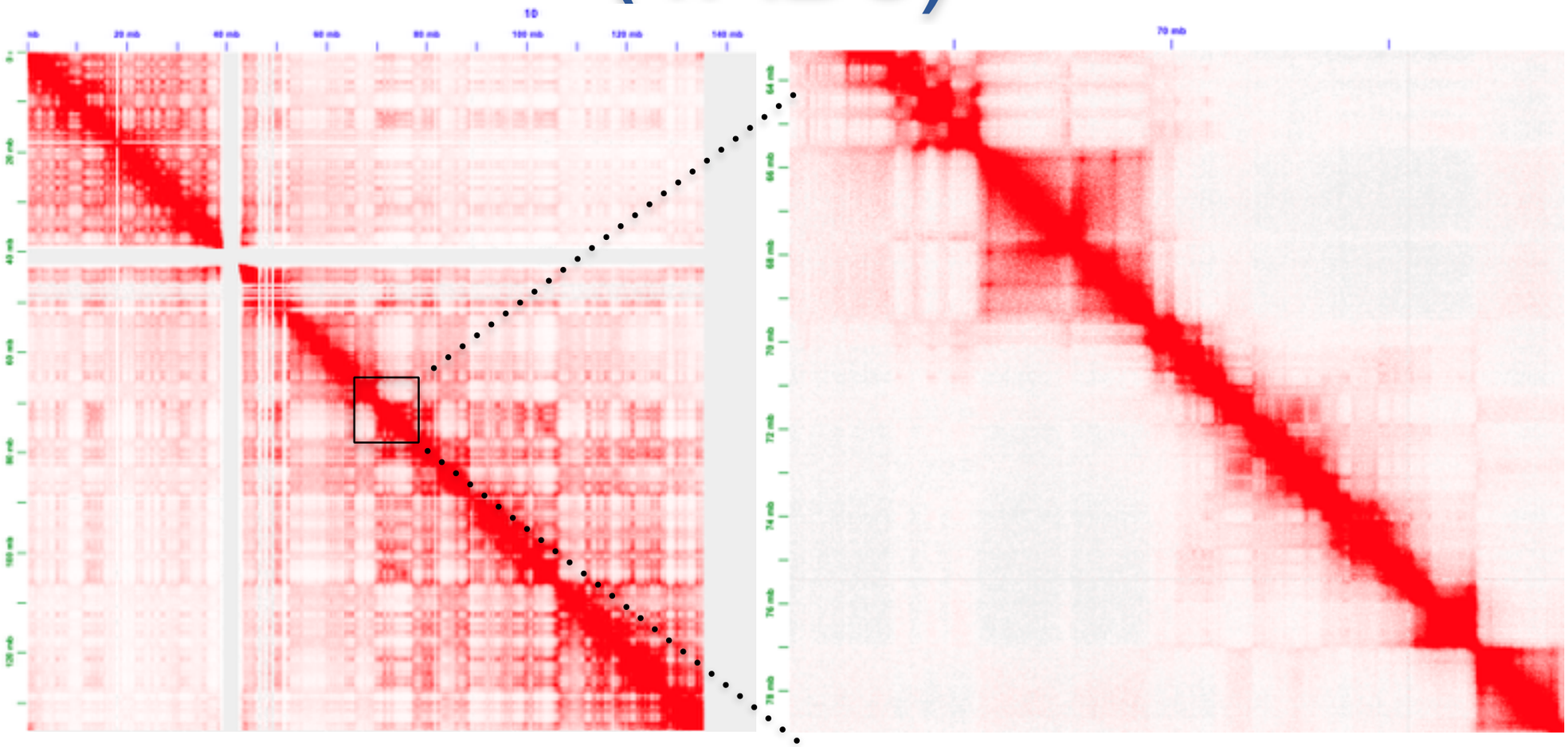
Topologically associating domains (TADs)



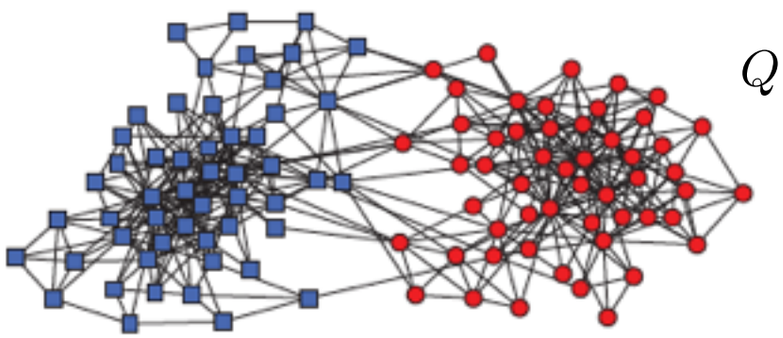
hESC chr1

Weinreb and Raphael, Bioinformatics 2015

Topologically associating domains (TADs)

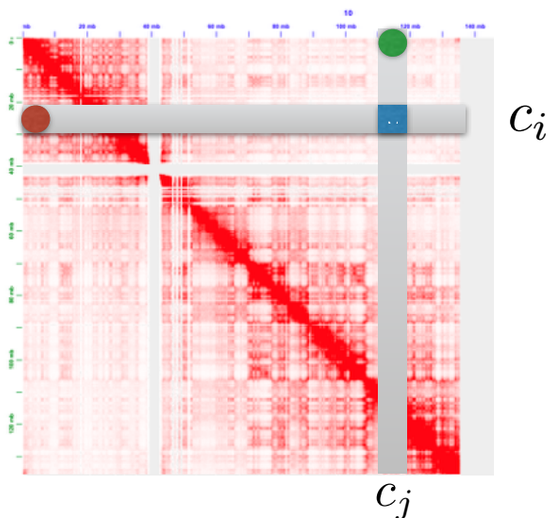


How could we identify such domains in multiple resolution?
Domains resemble network modules

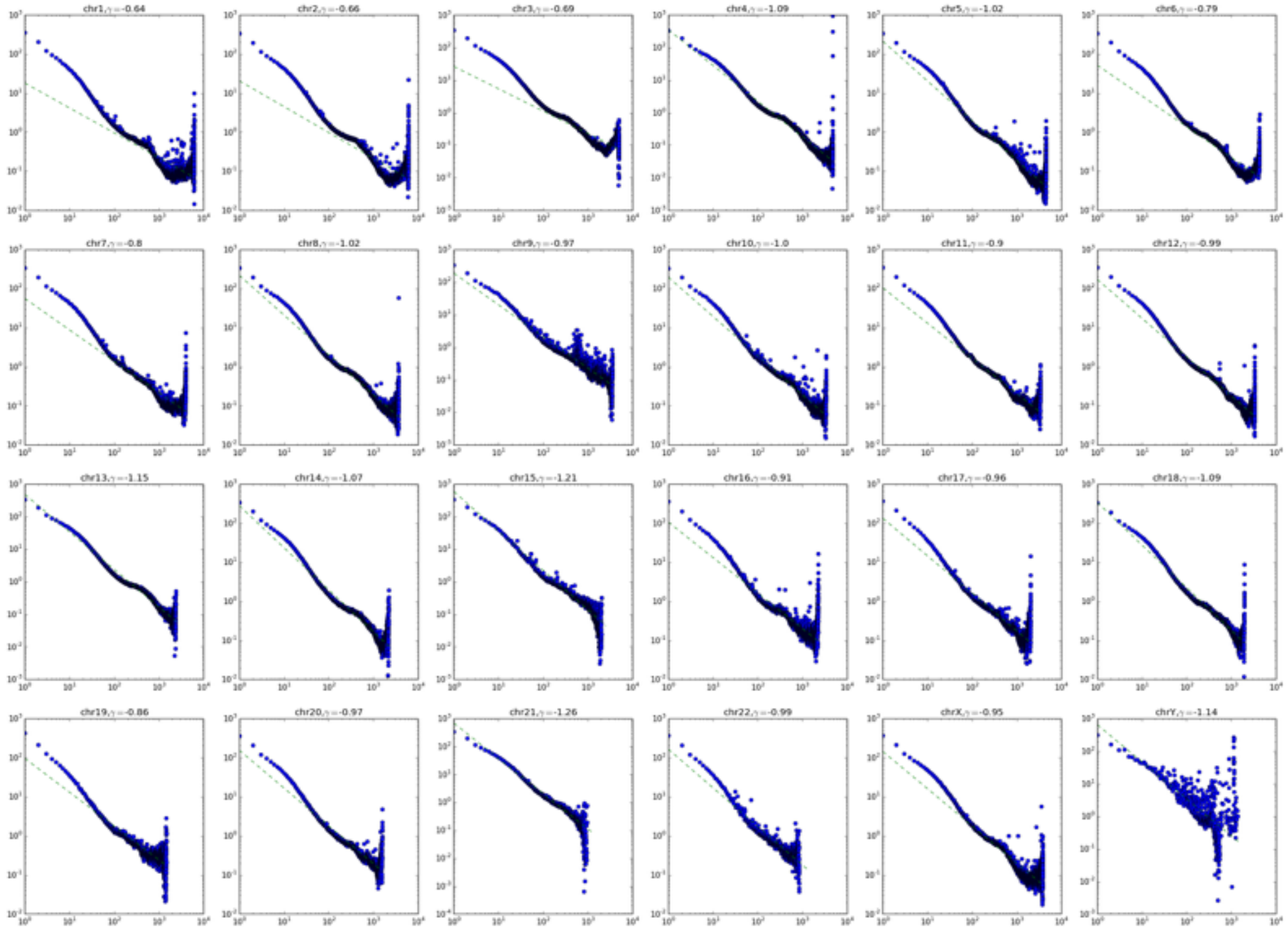


$$Q = \frac{1}{2m} \sum_{i,j} \left(W_{ij} - \frac{k_i k_j}{2m} \right) \delta_{\sigma_i \sigma_j}$$

$$Q = \frac{1}{2N} \sum_{i,j} (W_{ij} - \frac{c_i c_j}{2N}) \delta_{\sigma_i \sigma_j}$$

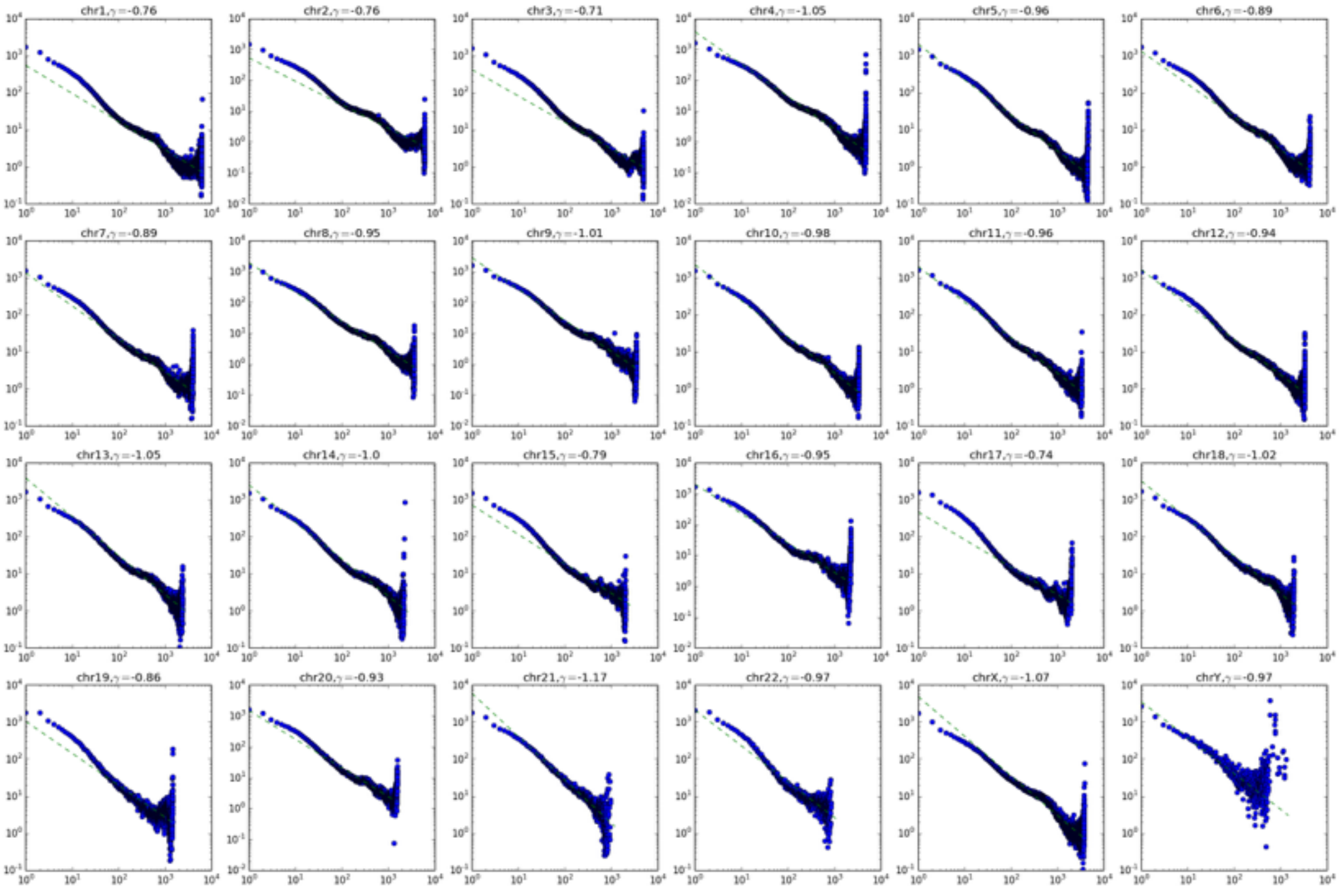


Number of contacts versus genomic distance



Data: Dixon et al. 2012

Number of contacts versus genomic distance



Data: Dekker A549C

Identifying TADs in multiple resolutions

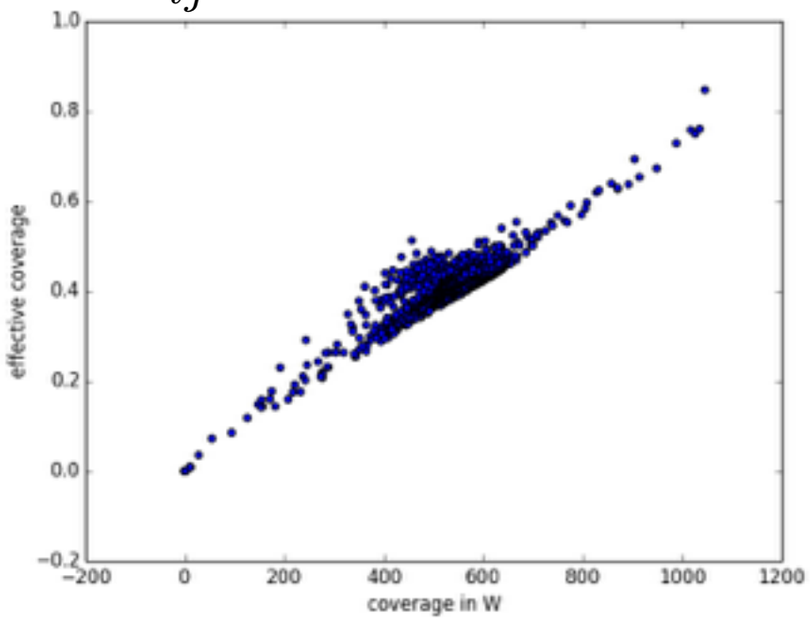
A null model that takes into account of the genomics distance

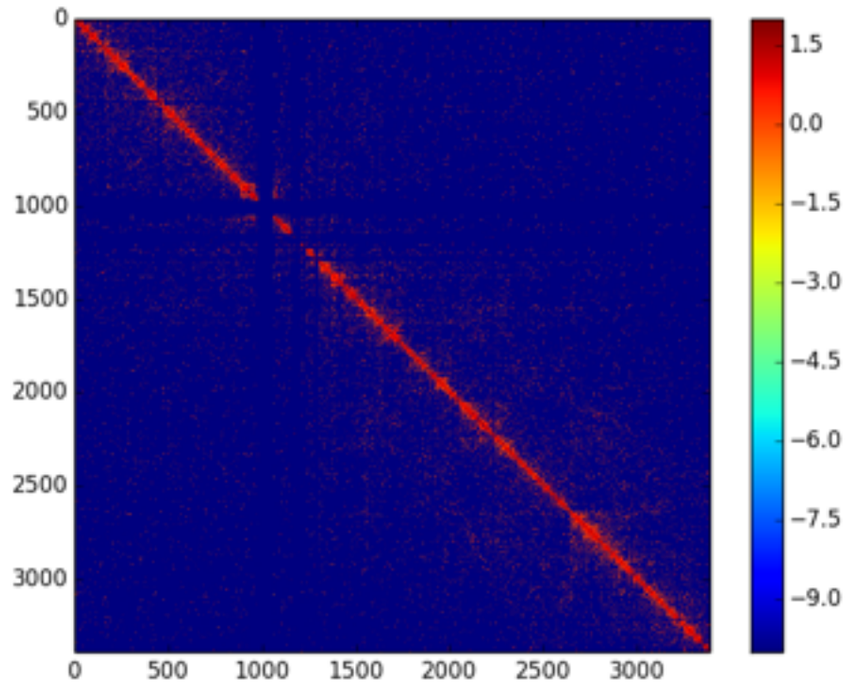
$$E_{ij} = c_i^* c_j^* f(|i-j|)$$

contraints

$$\sum_j E_{ij} = \sum_j W_{ij} = c_i \quad \text{coverage of loci i}$$

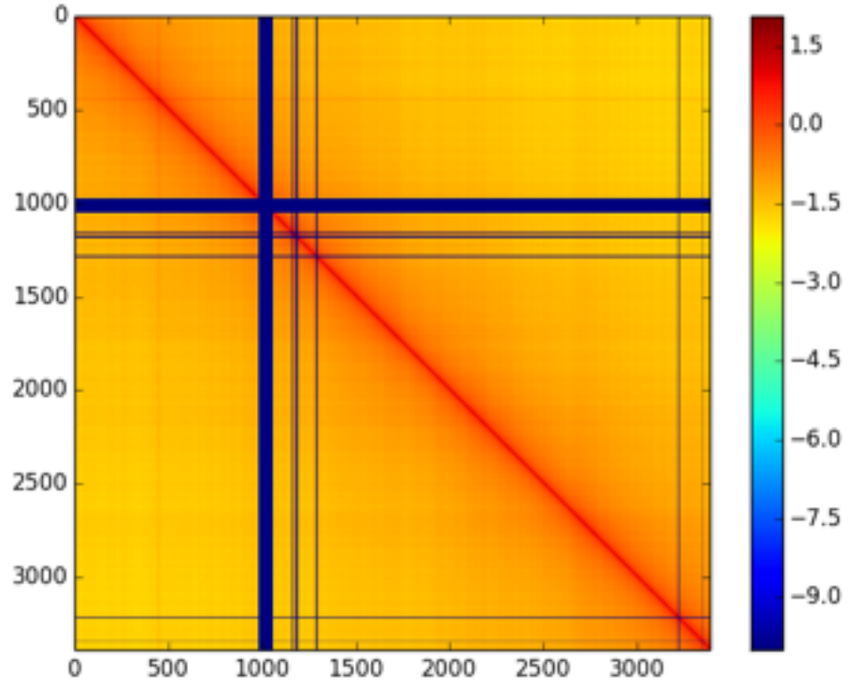
$$\sum_{ij} E_{ij} = \sum_{ij} W_{ij} = 2N \quad \text{total number of reads}$$





W

chr10. (40kb bins)

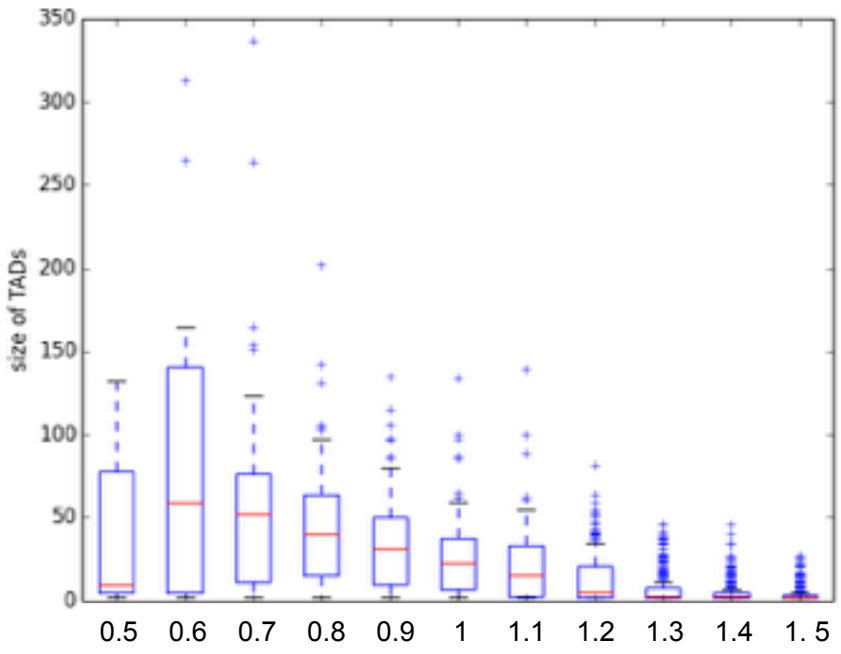


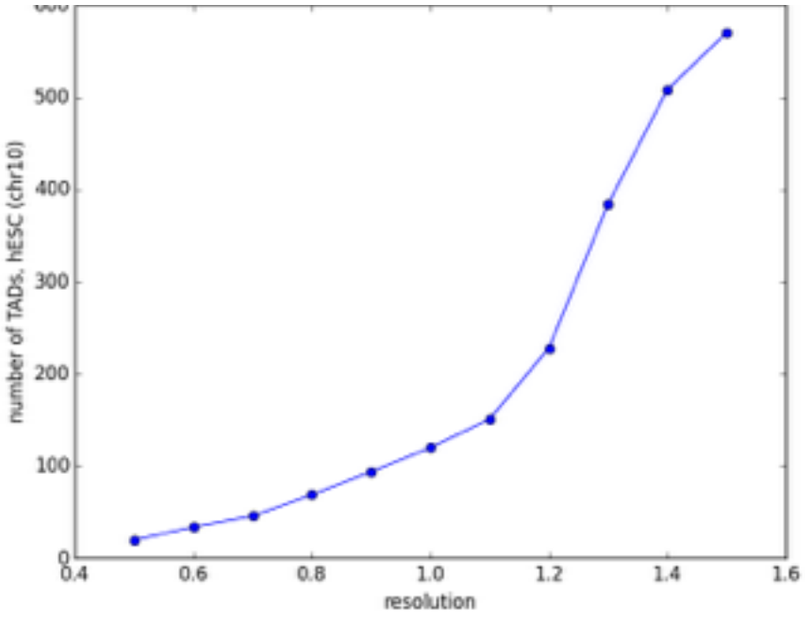
E

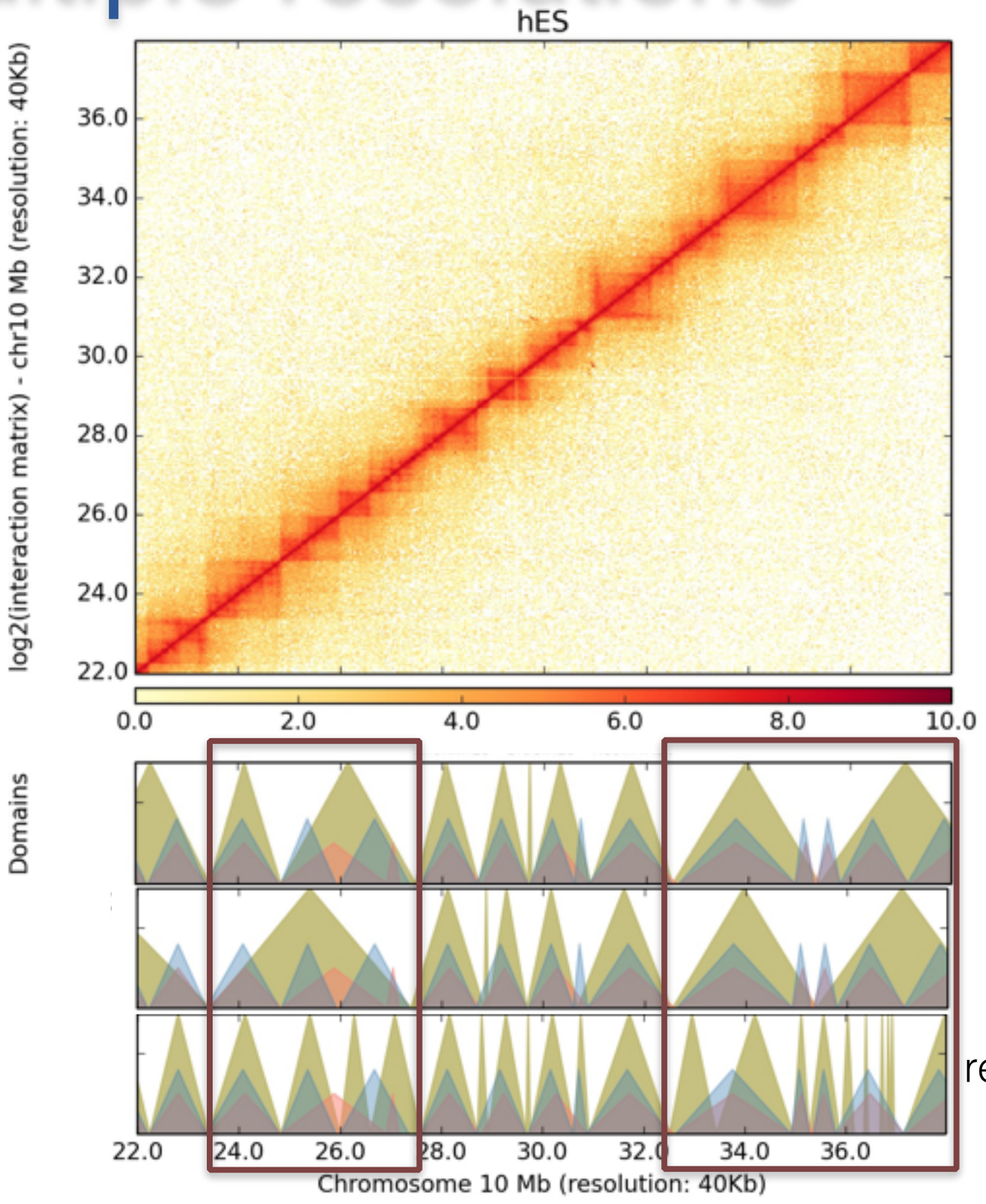
MrTAD Finder: a tool to identify TADs in multiple resolutions

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

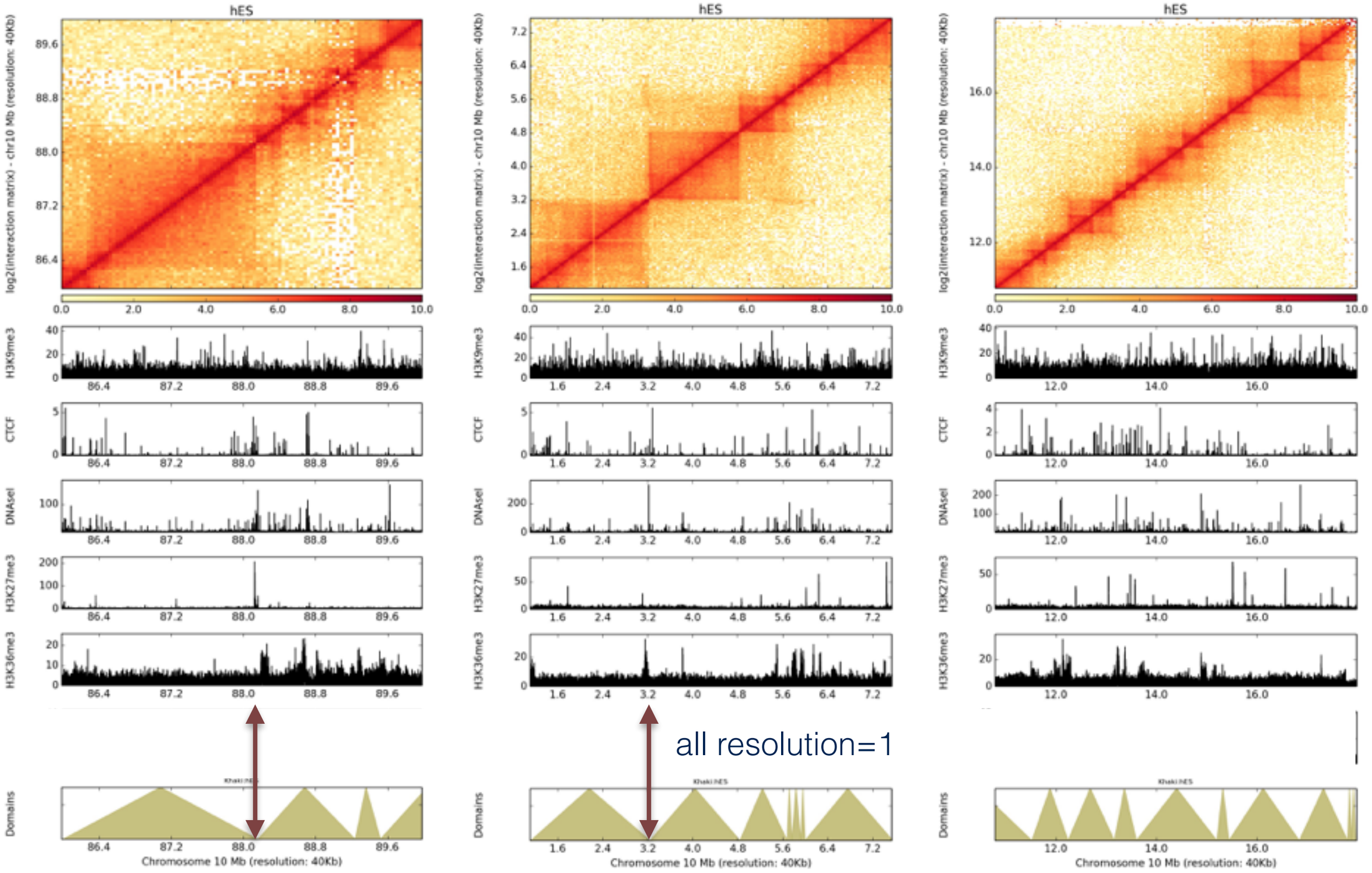
resolution parameter



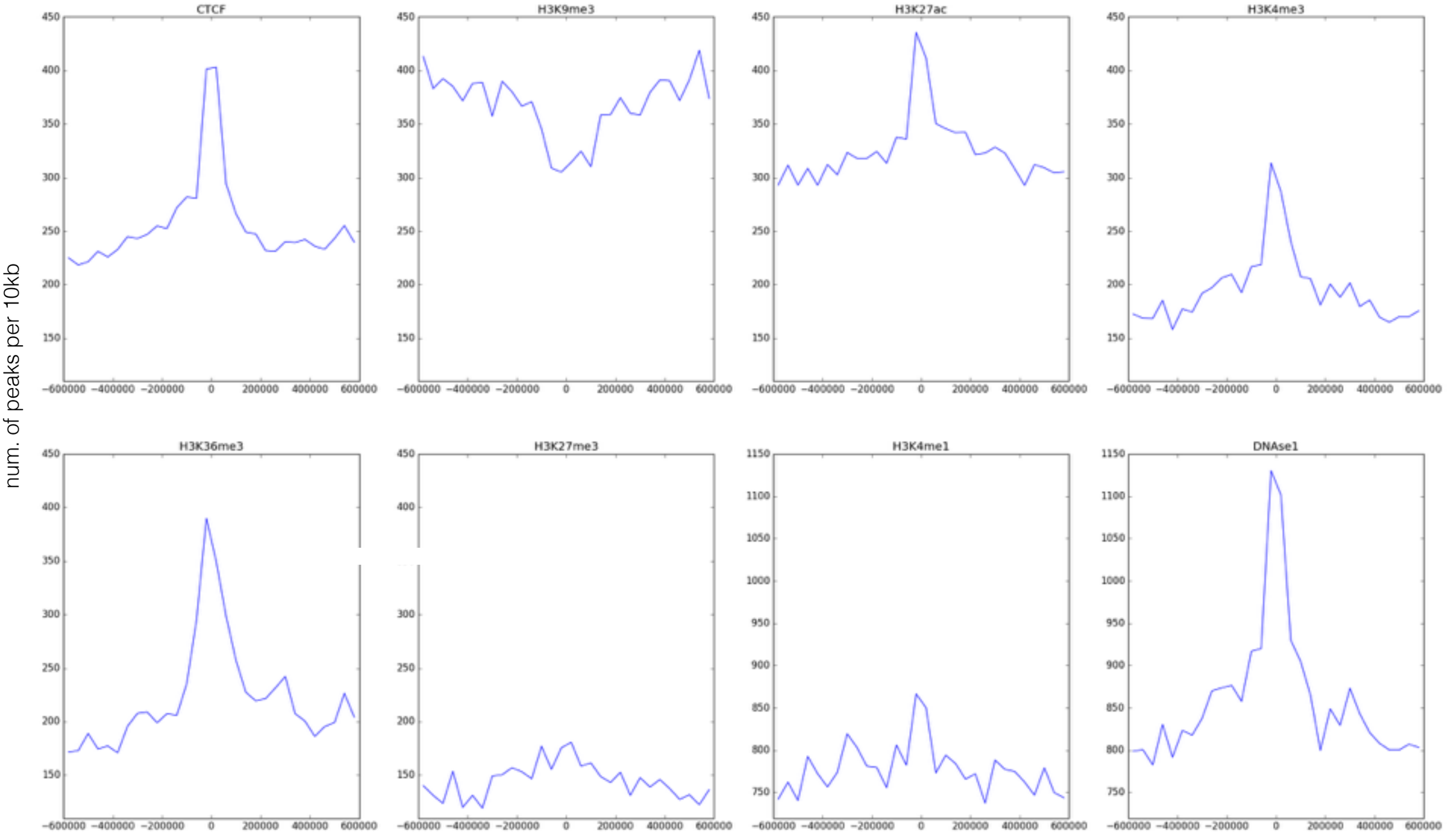




Relationship between TADs and chromatin features

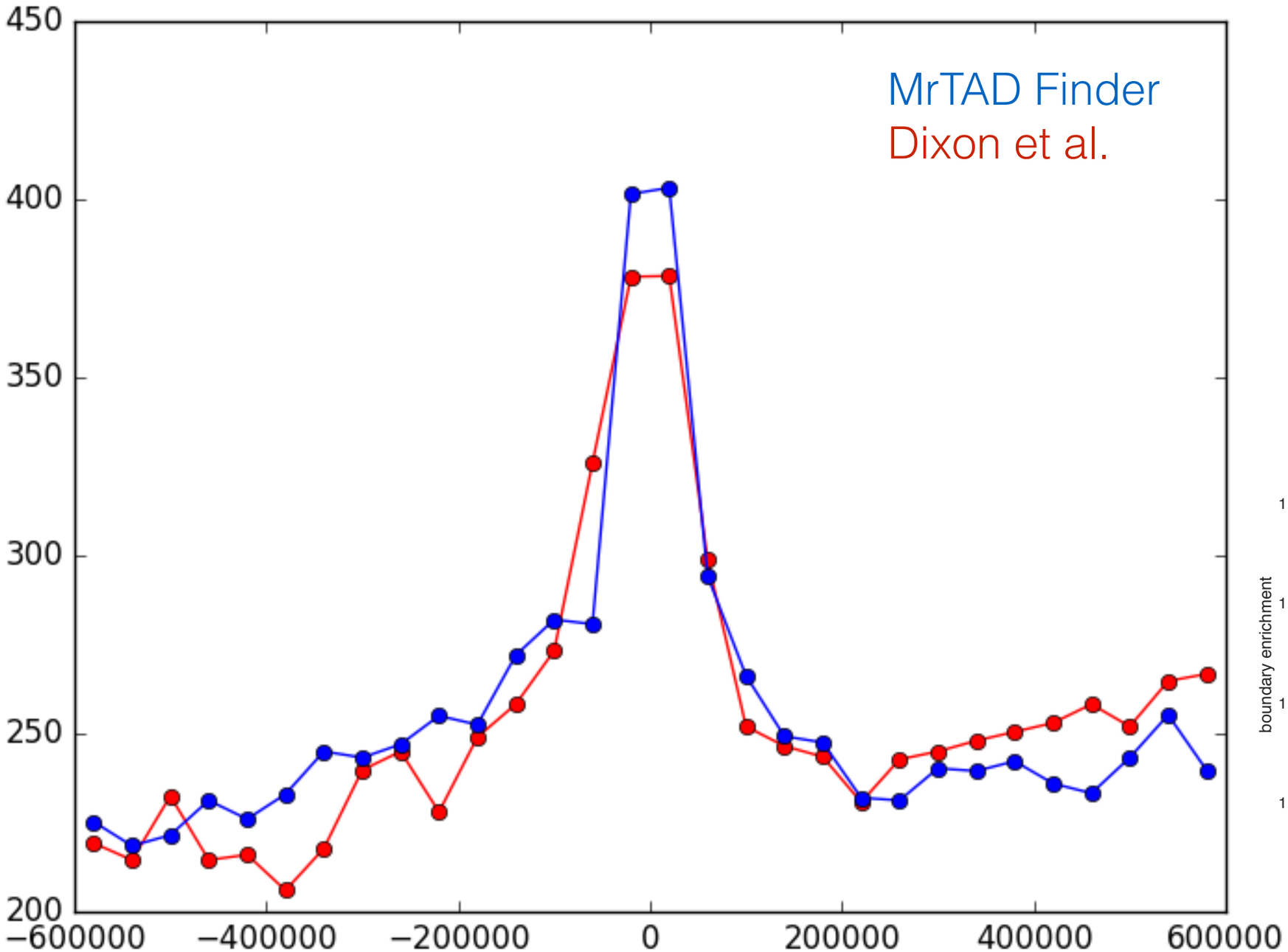


chromatin features near domain boundary



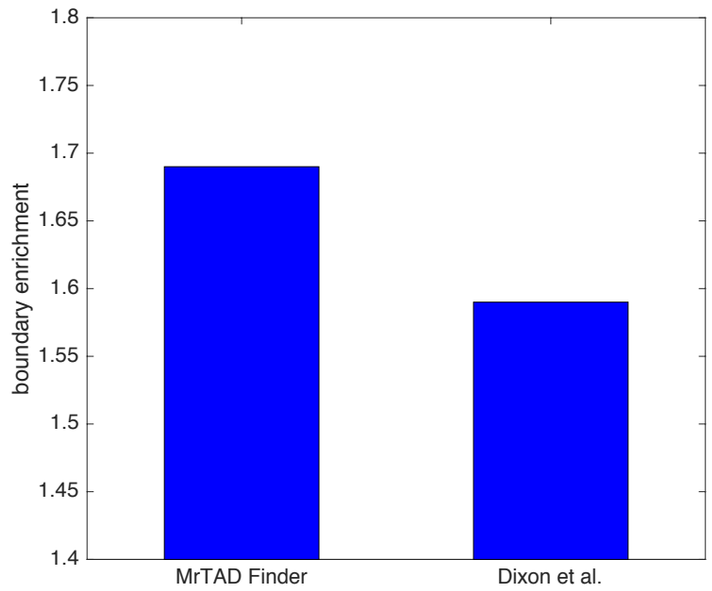
res=1, hESC, all chr

Comparison with existing algorithms

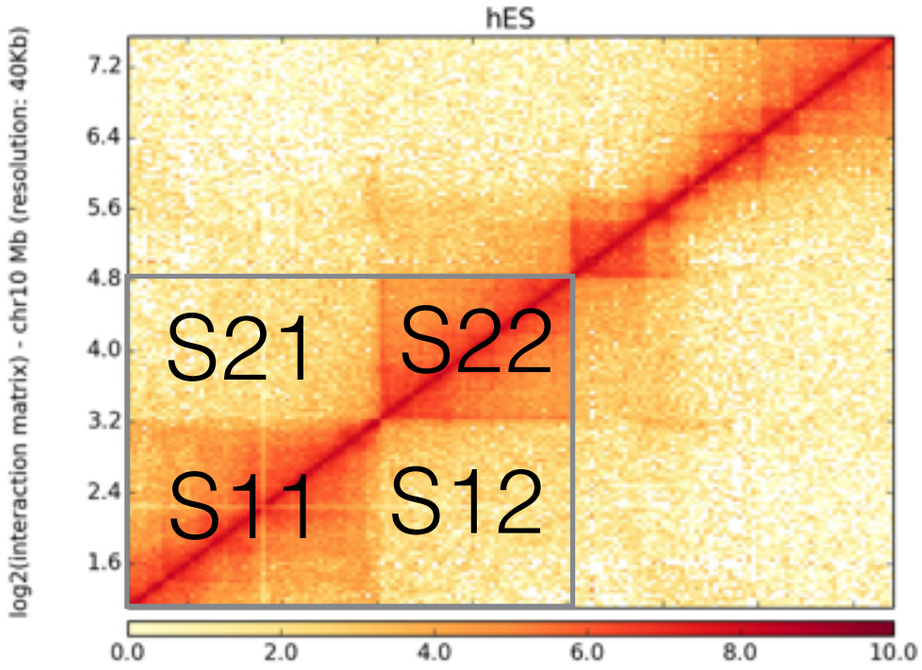


res=1, hESC, all chr

res=1

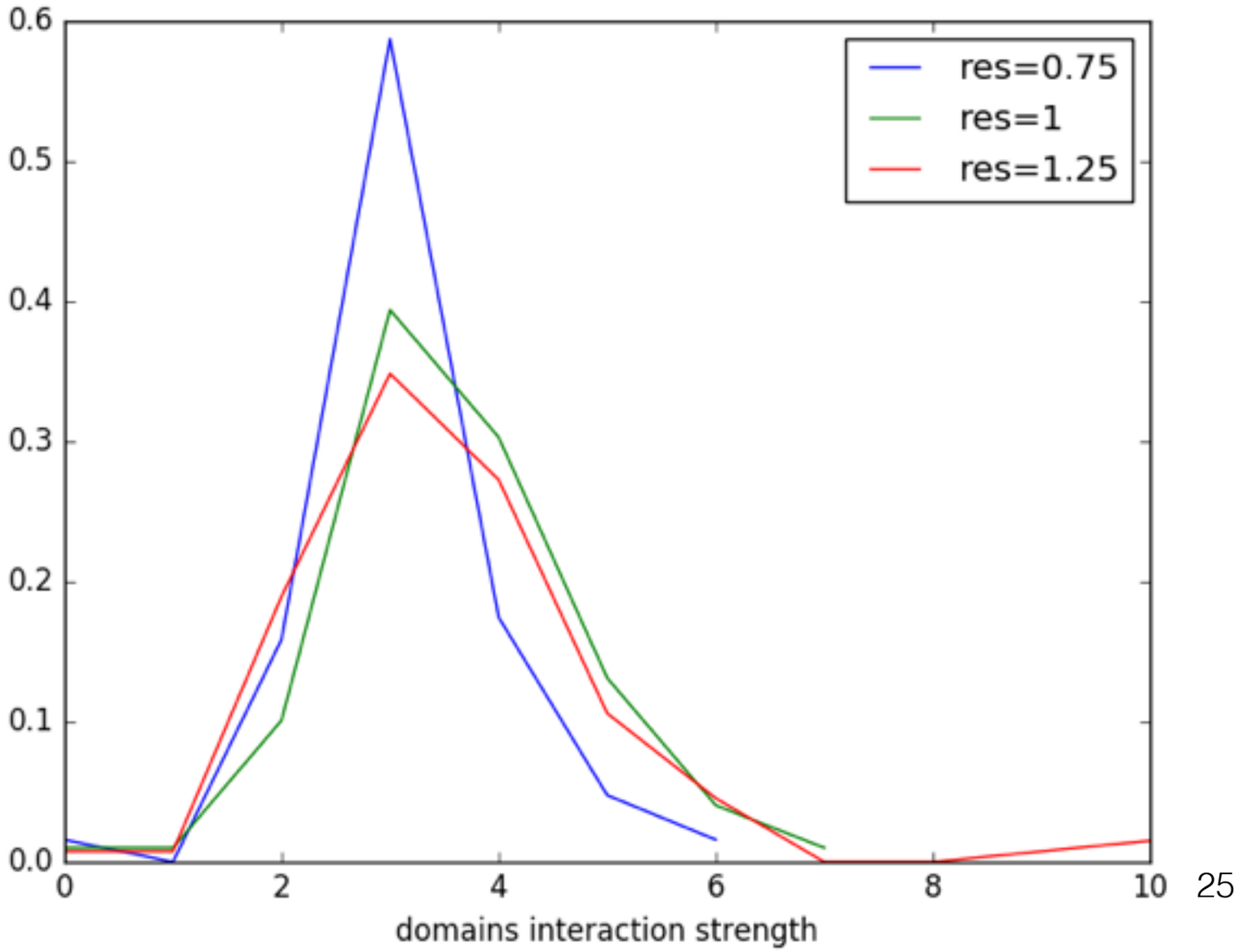


Domains interaction strength



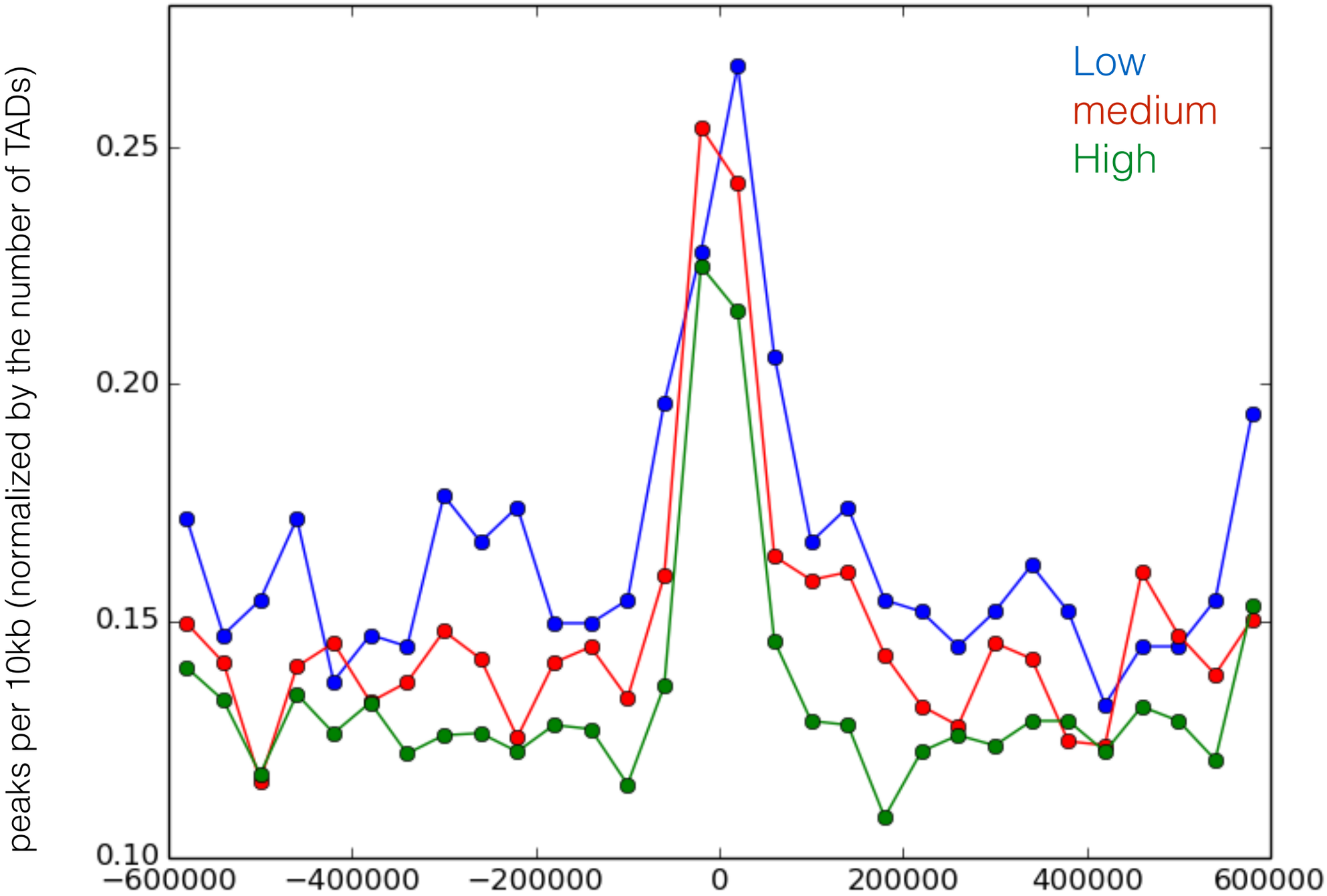
interaction strength between 1 and 2 = $$\frac{S11 + S22}{S12 + S21}$$

$$S_{mn} = \sum_{i \in m, j \in n} |(W_{ij} - \gamma E_{ij})|$$



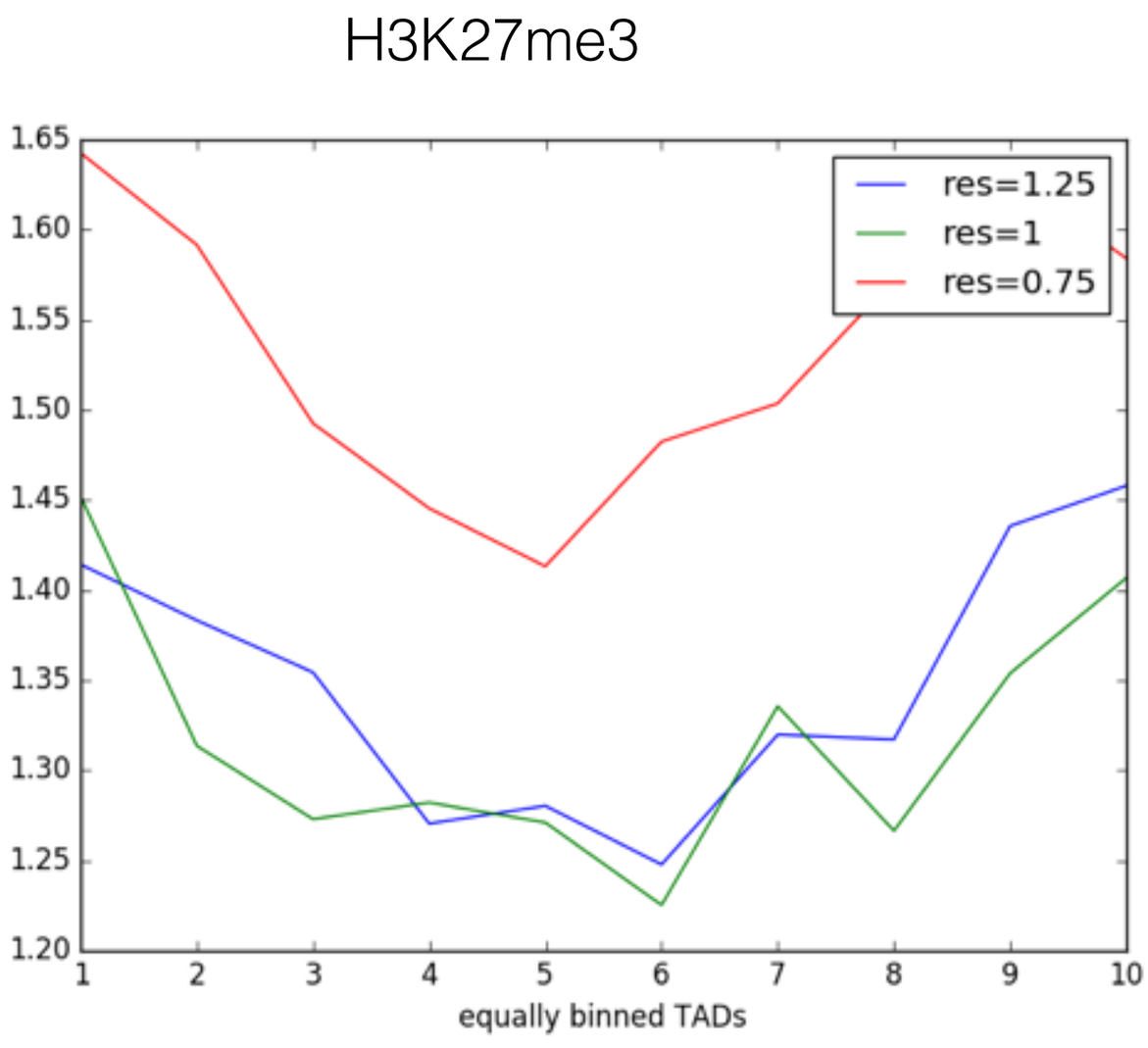
hESC, chr 1

Boundary features with respect to resolution



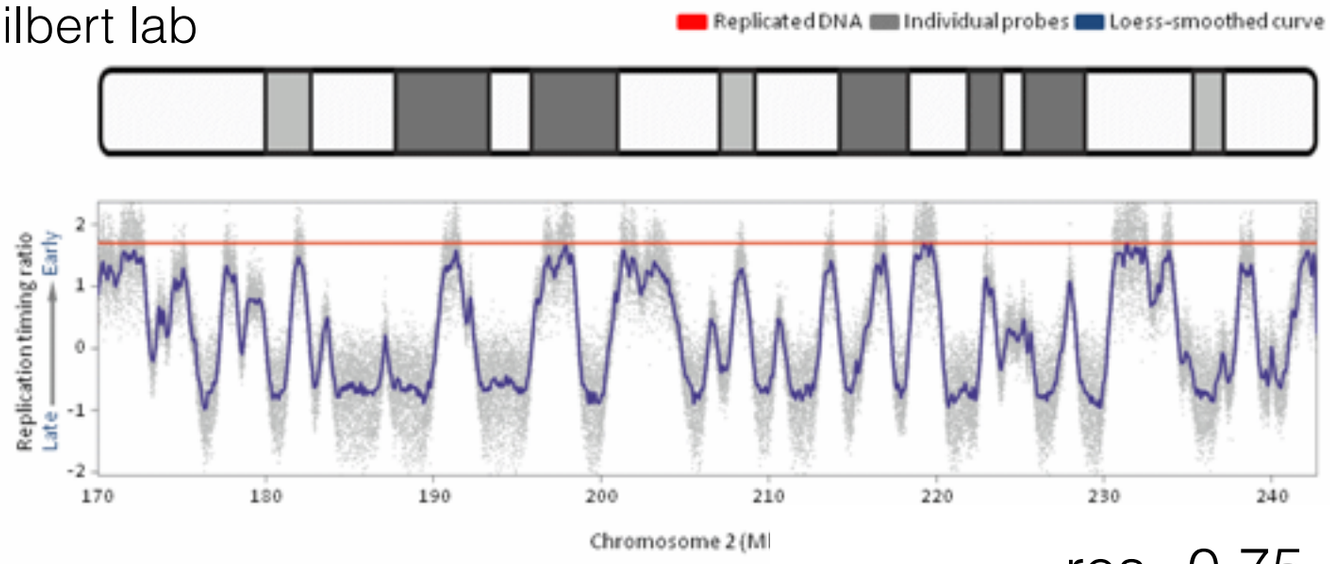
Chromatin signatures in different resolution

- various chromatin features, where are they distributed along the domains?
- effects of the resolutions? types of domains?
- enrichment of peaks/signals?



Replication timing

Gilbert lab

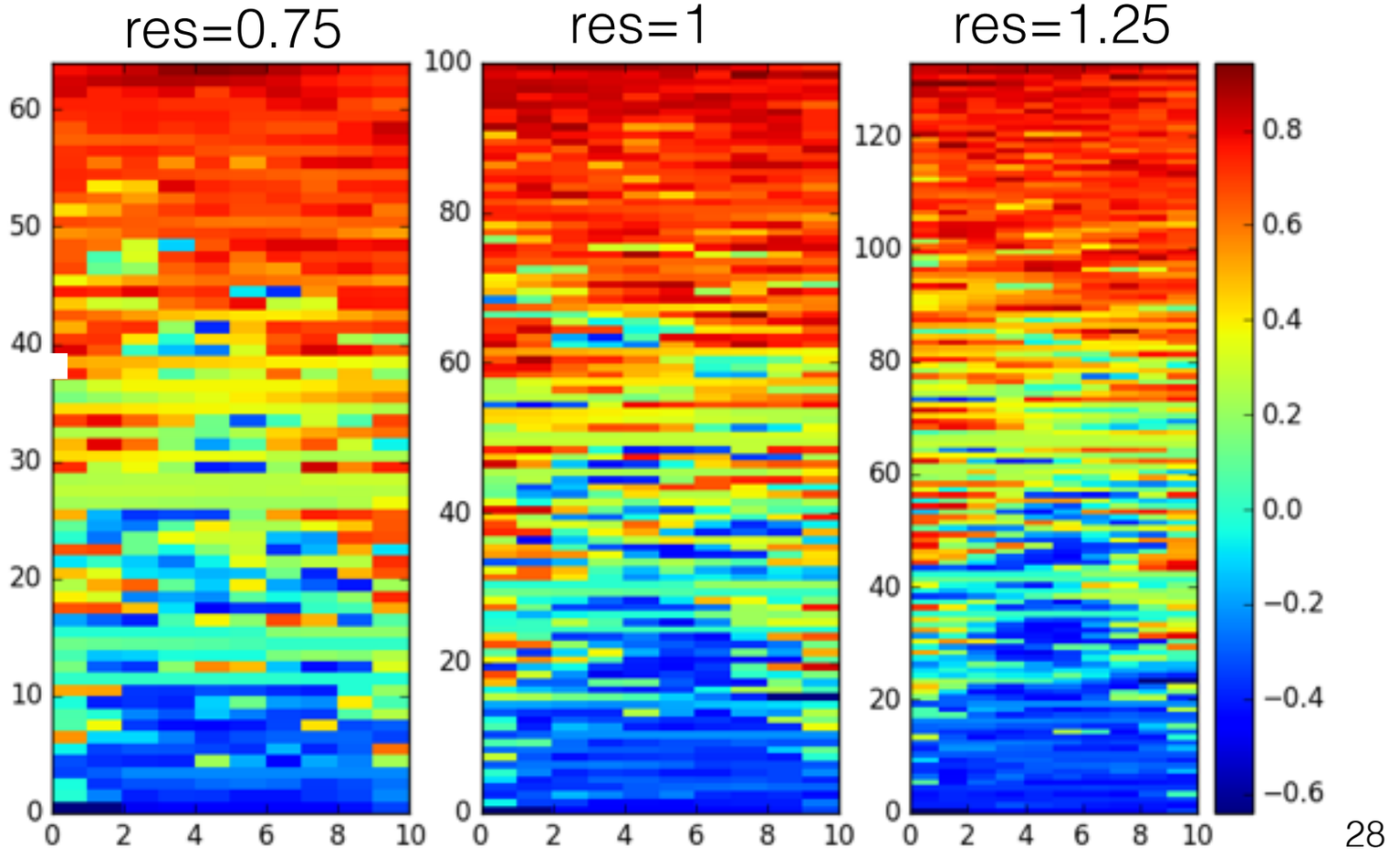


reply-chip data:
H1 ESC

early replication domains

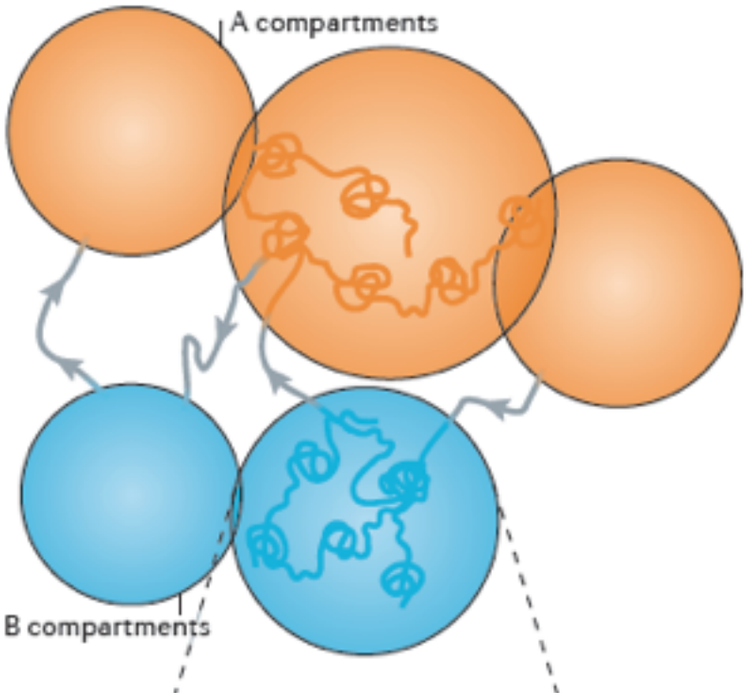
U-shaped domains

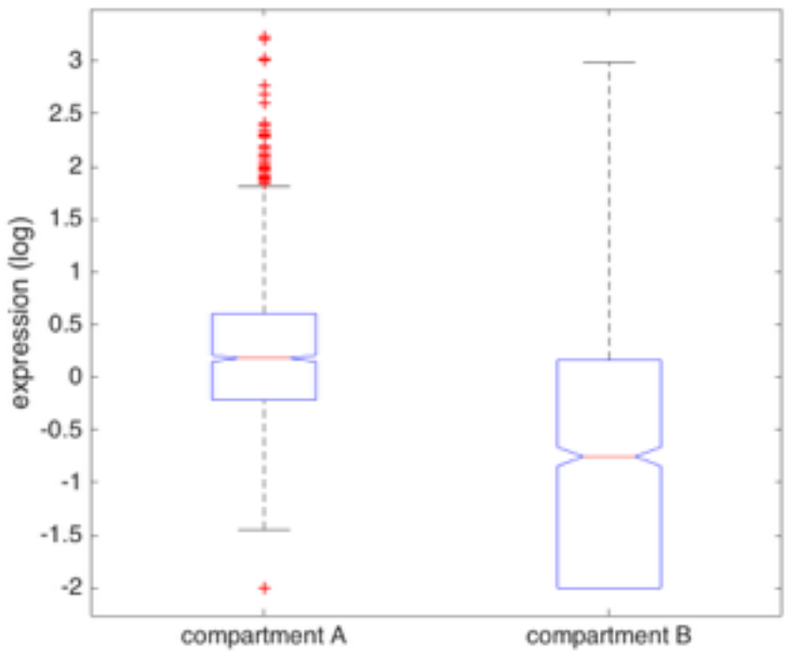
late replication domains

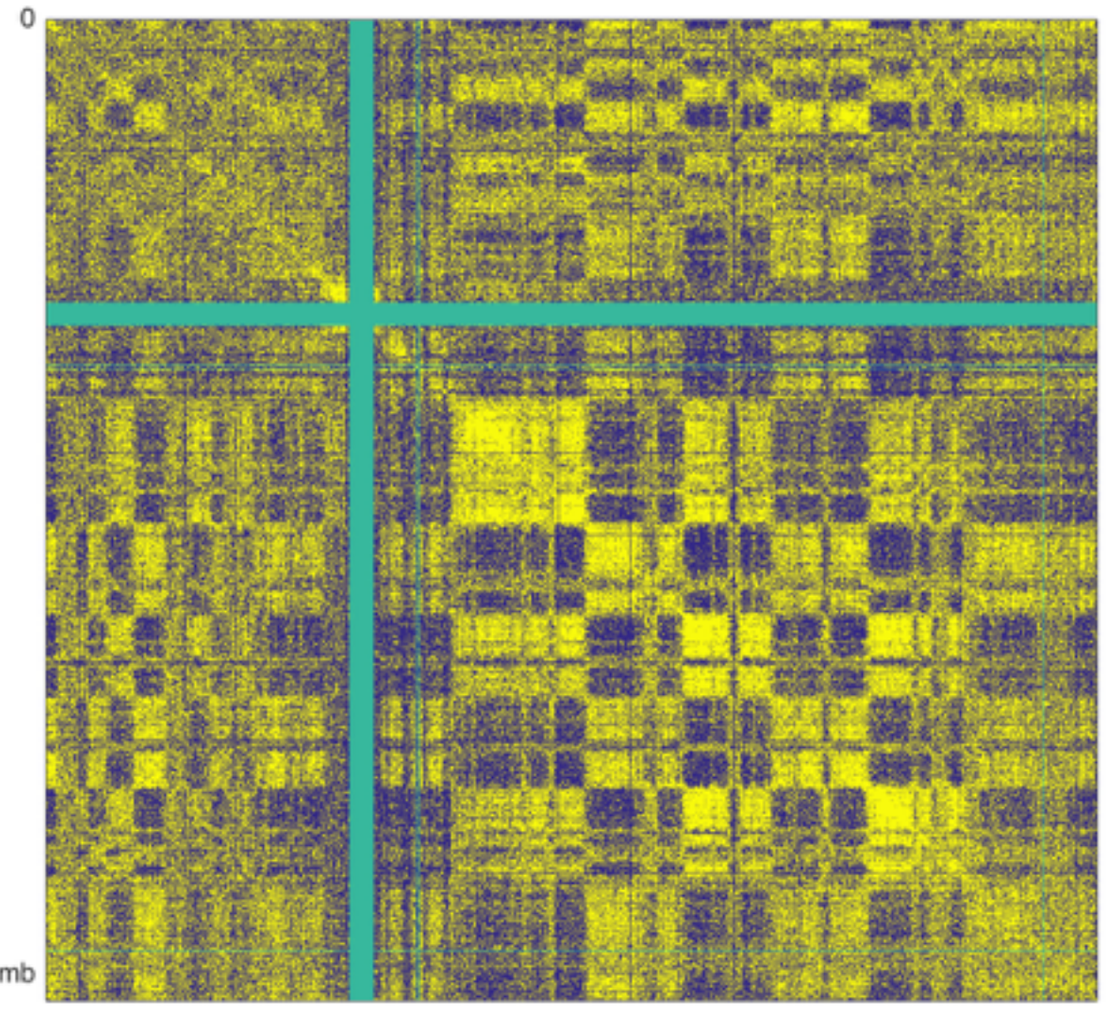


Compartments versus domains

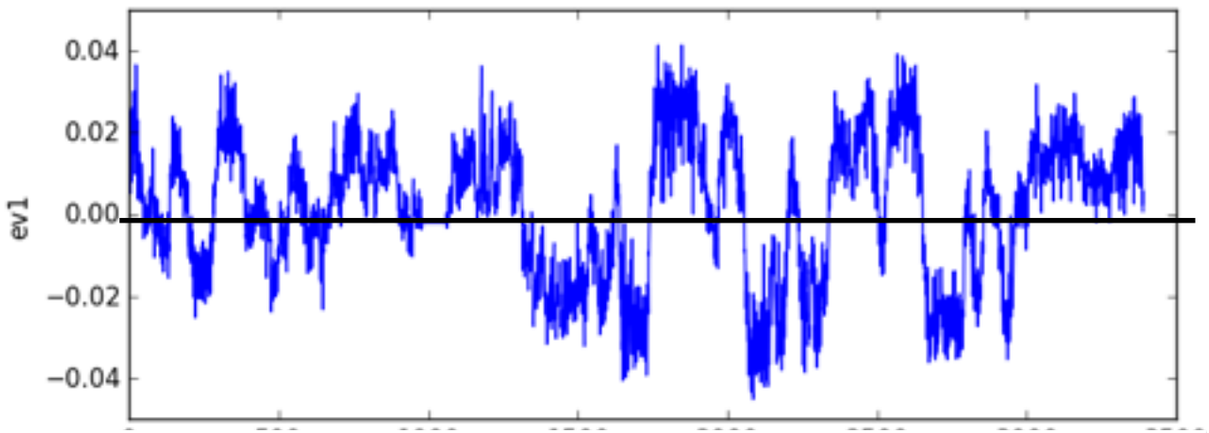
$$C_{ij} = cor(W_{ij}/E_{ij})$$



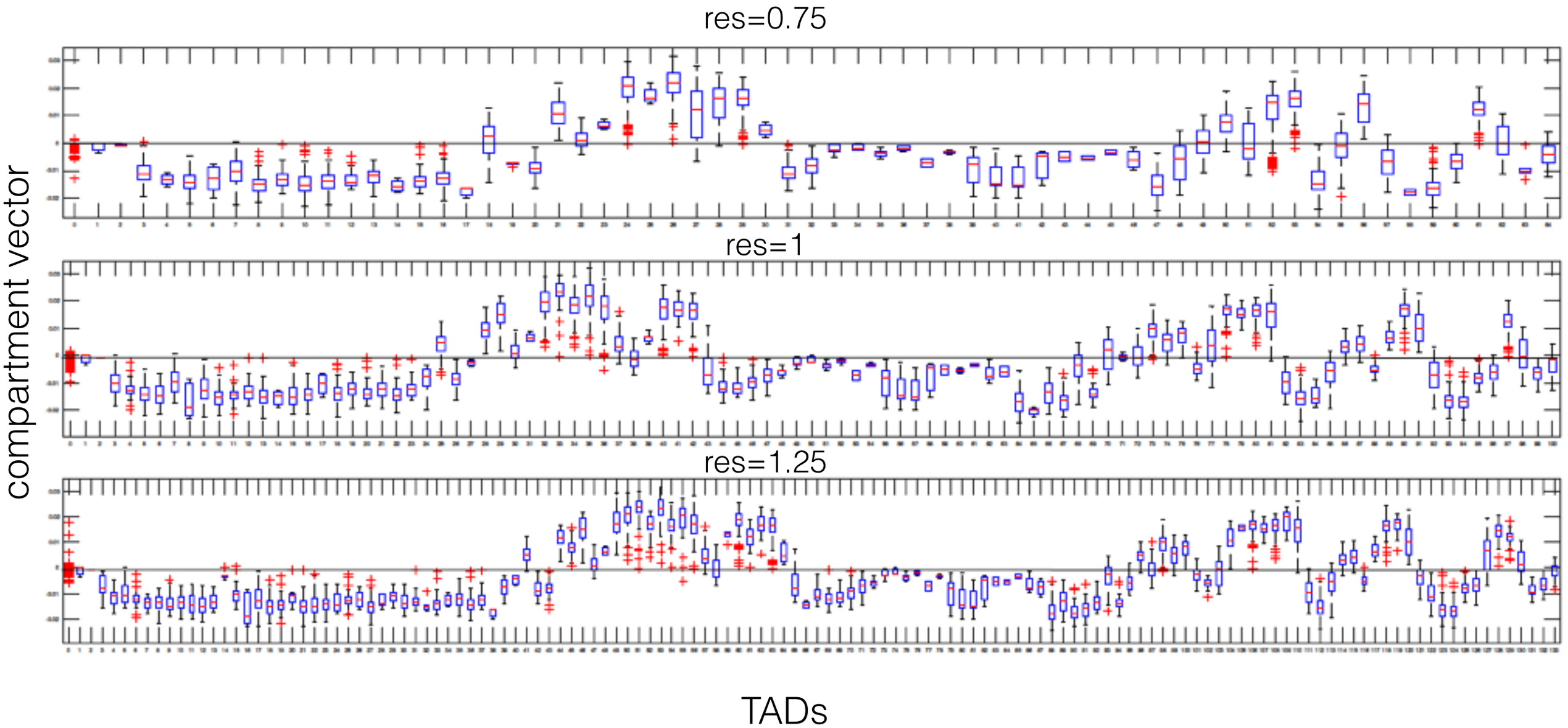




hESC chr10



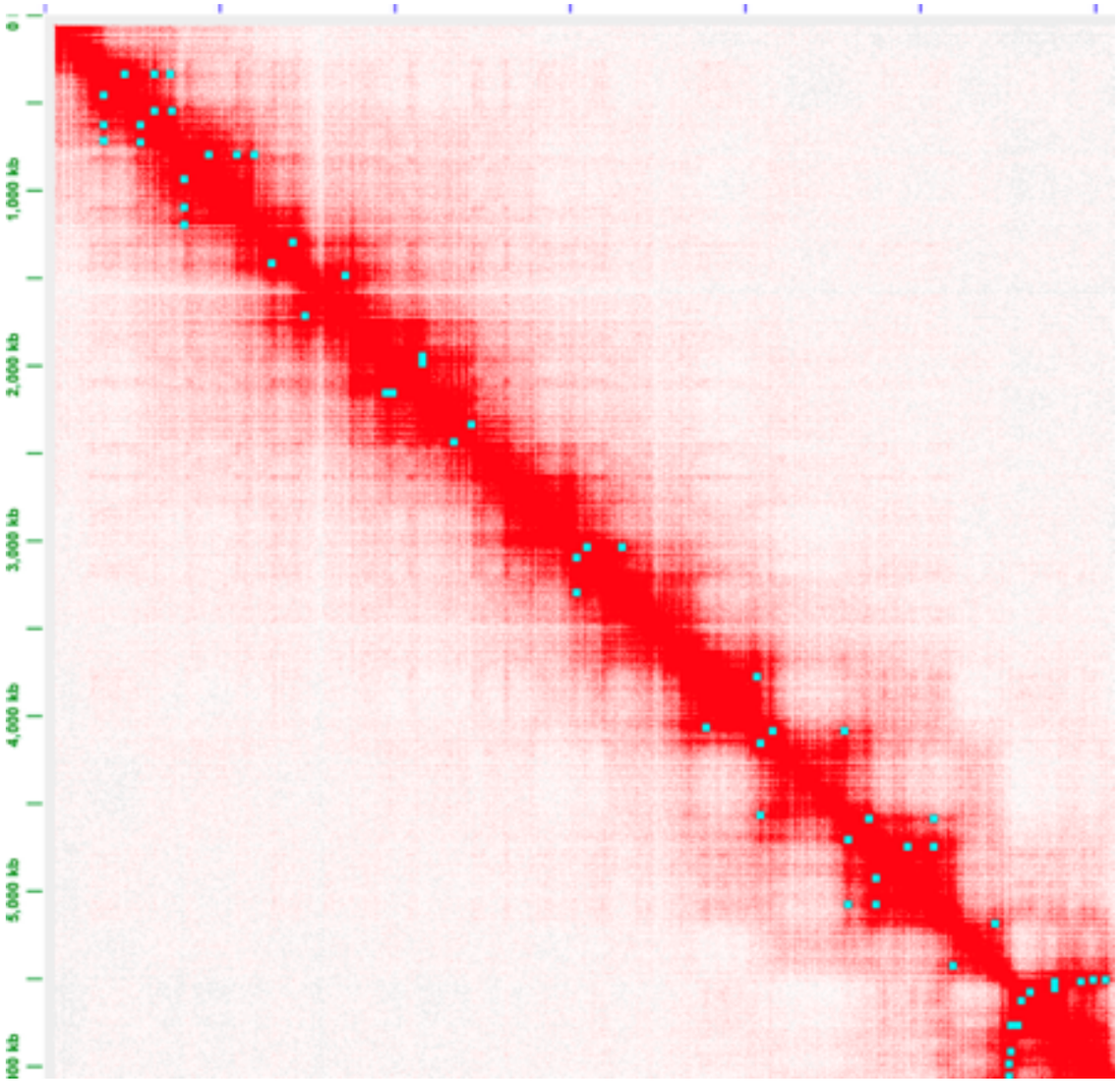
Compartments versus domains



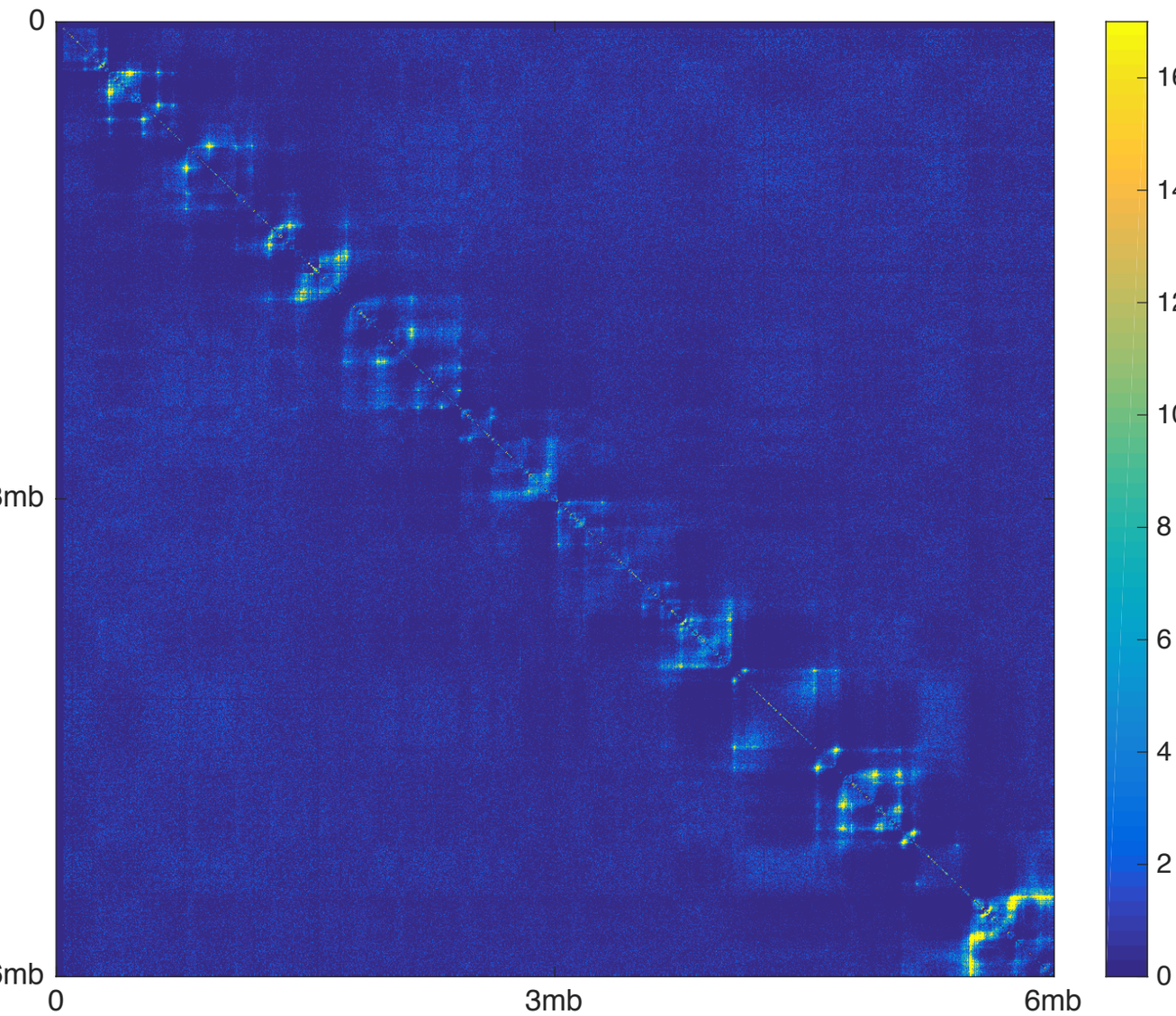
hESC chr1

Enrichment of contacts

GM12878. chr 10. bin size=5kb



cf. real contacts vs. null
assume Poisson distribution (log P)



Likely to have many false positive because of over dispersion.

Data:Rao et al. Cell 2014

Summary and the next steps

- MrTAD Finder: a novel tool to identify TADs
  - take into account of a background that captures genomic distance (the idea could be used in other network context)
  - based on global optimization inspired by network modules as oppose to local approaches in existing methods
  - with a concept of continuous resolution, more general than a bottom-up hierarchical structure

Summary and the next steps

- Certain characteristic features for different resolutions
  - histone marks
  - expression (active and inactive domains)
  - replication timing (domain with multiple resolutions too?)
  - k-mer frequency in TADs across multiple resolutions (with ANS, Yunsi)
- Can interaction strength be reflected by chromatin features (combinatorially) near the boundary? CTCF orientation?
- 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)

Acknowledgement

- Tech
  - ANS, TG, JR, RK, AH, MG