messing around with Hi-C data

Koon-Kiu Yan

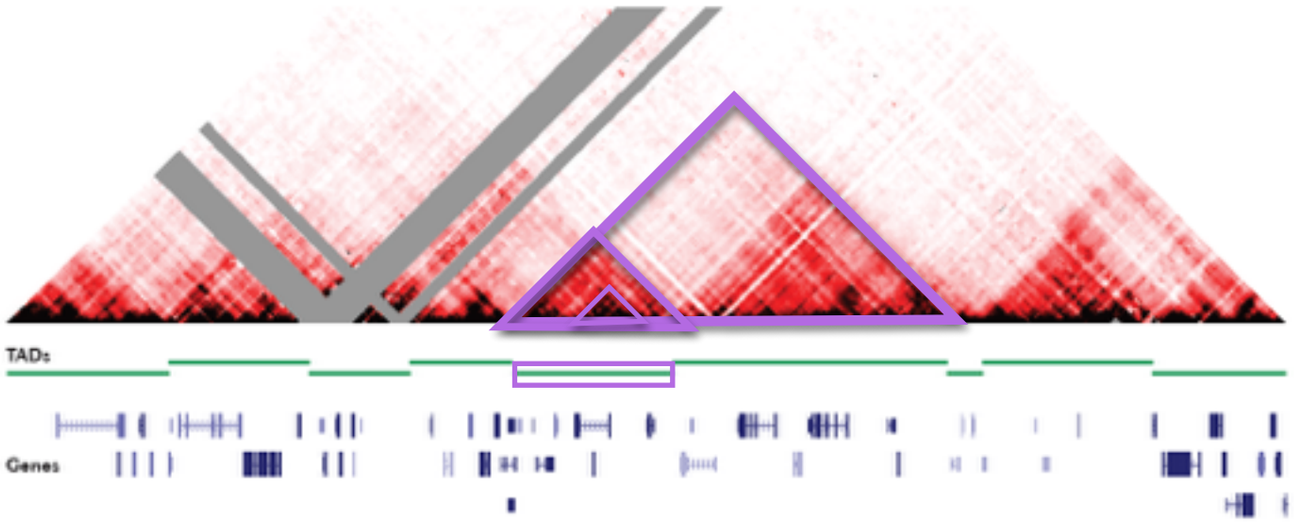
Jan 2015

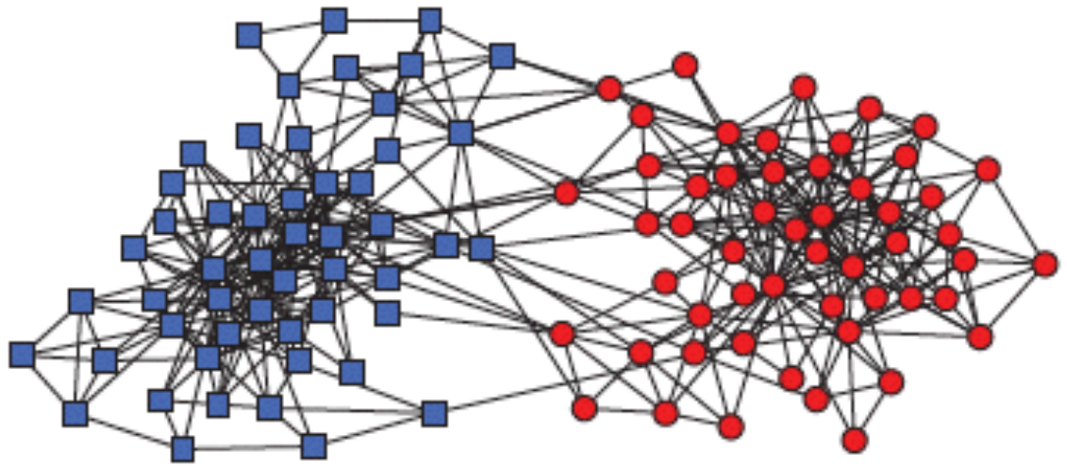
Several relevant updates

- Installed Dekker lab's pipeline. Re-processed a few early datasets.
- Look at some of the data from Aiden lab, Cell 2014
- Visualization tool: HiCPlotter, Juicebox

To identify topological domains based on network modularity detection

multiple resolutions





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

adjacency matrix

degree of i

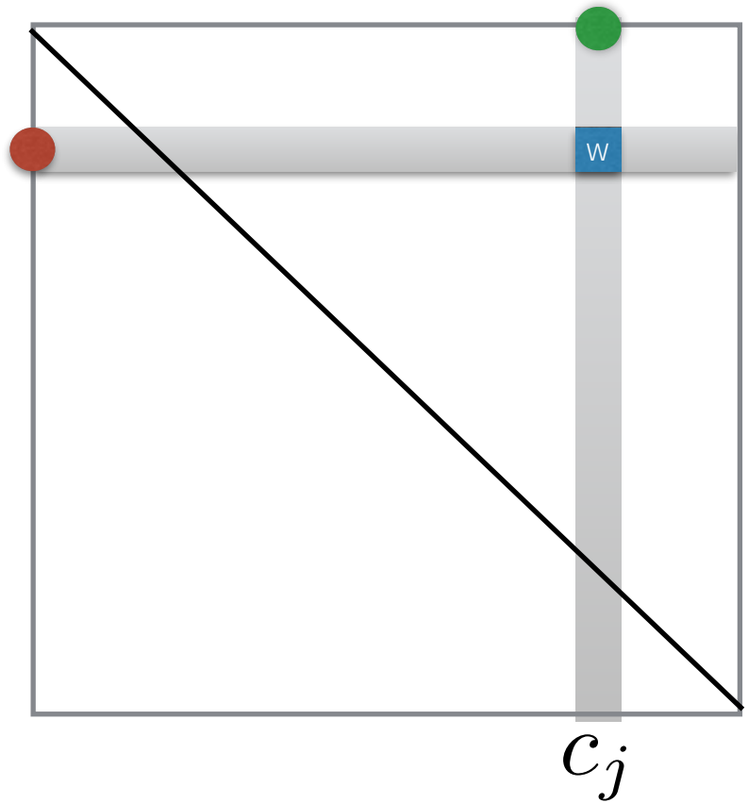
whether or not i, j are in the same module

number of edges

expected number of edges between i and j

Naive null model

Hi-C contact matrix



N: the total number of reads

relative coverage of loci i= $\frac{c_i}{2N}$

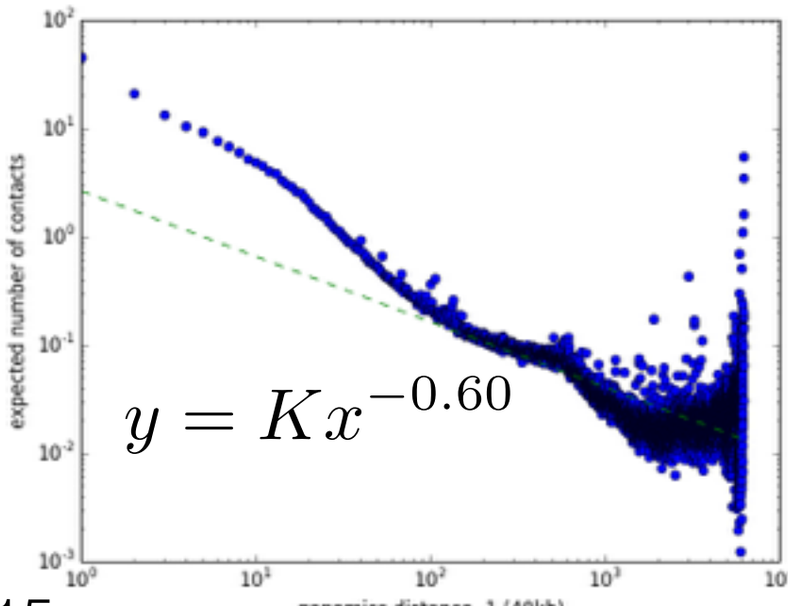
expected number of reads between i and j $= c_j \frac{c_i}{2N}$

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

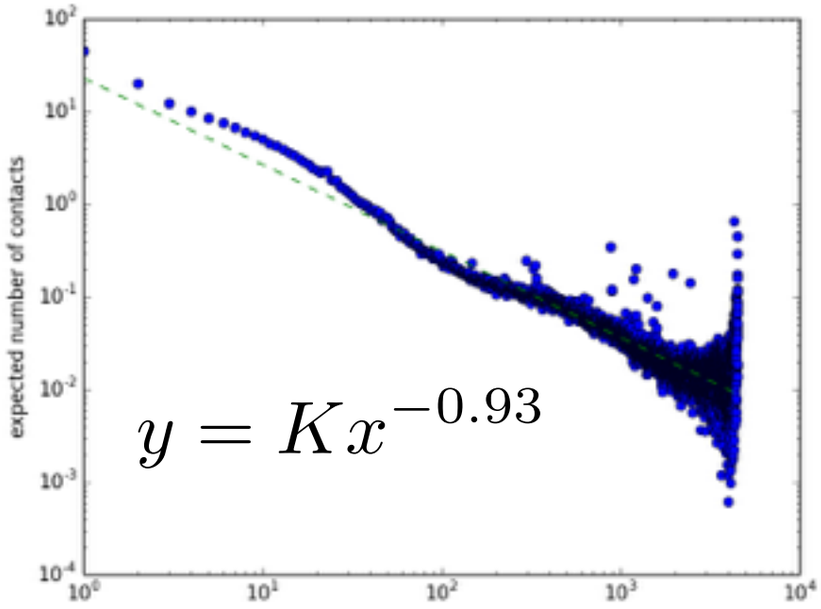
Does not take into account the genomic distance between i and j

Number of contacts vs genomics distance

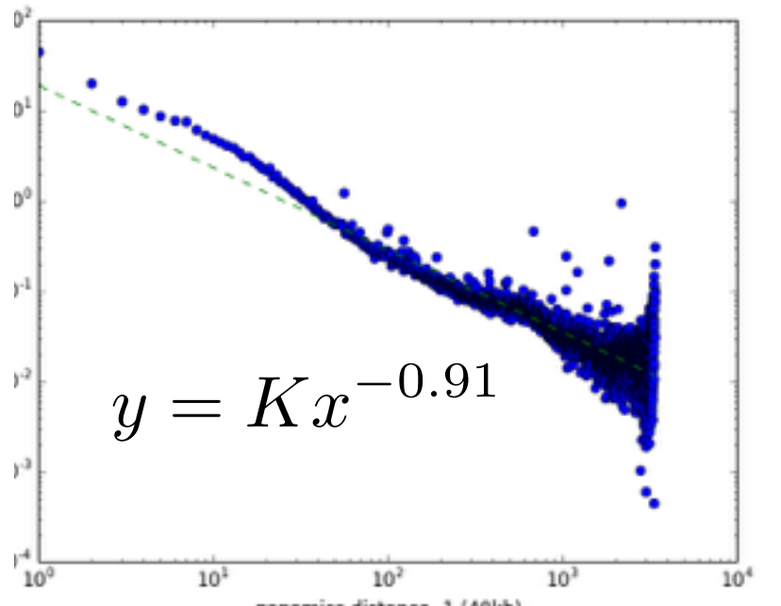
ch1



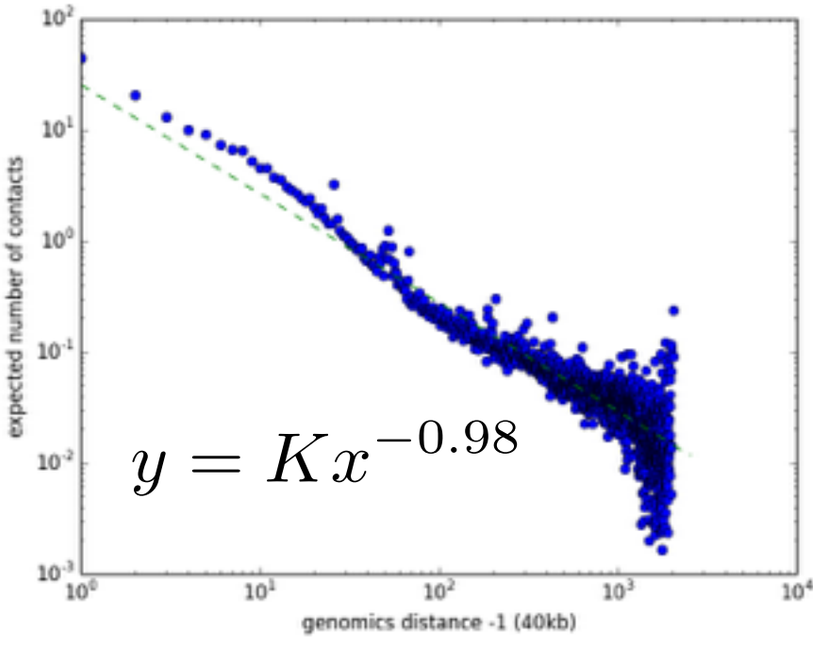
ch5



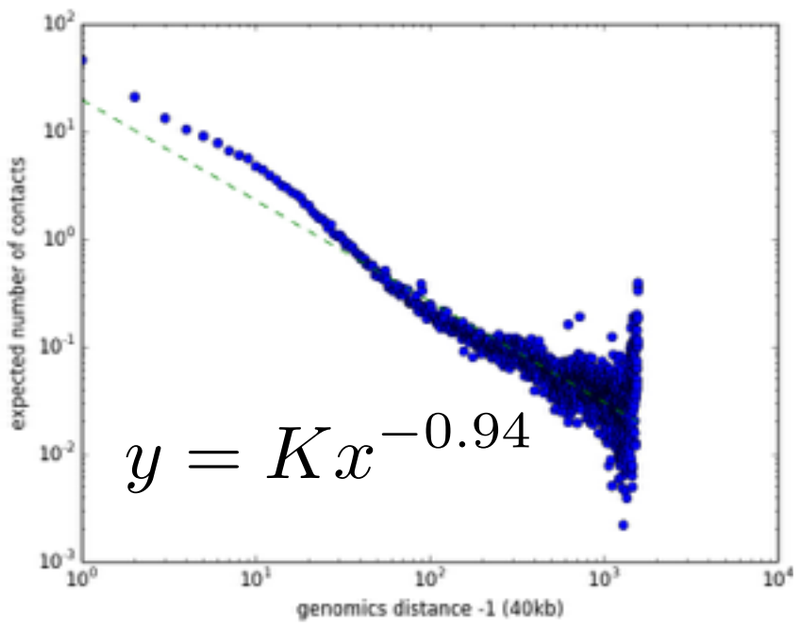
ch10 Dixon et al. 2012 hES



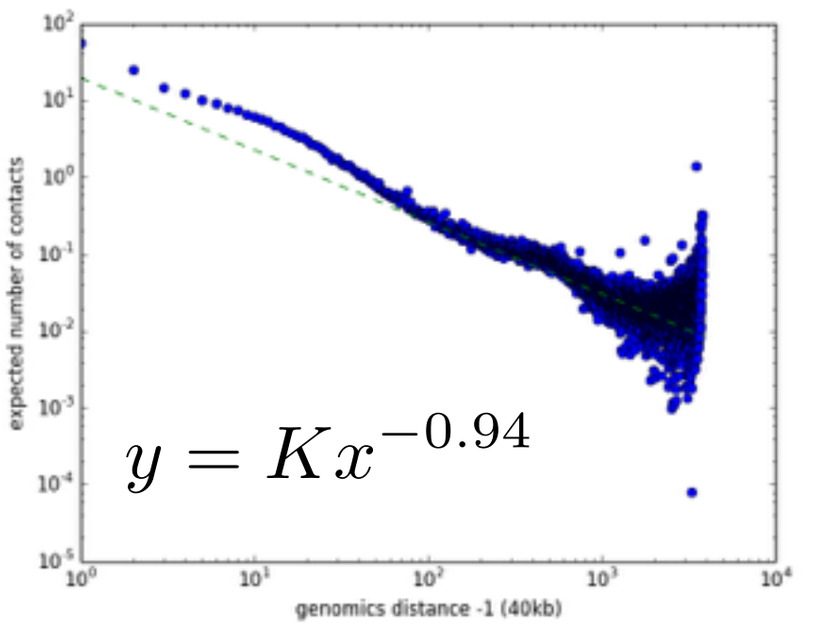
ch15



ch20



chX



A null that takes into account of genomic distance

naive

$$E_{ij} = c_i^* c_j^*$$

$$c_i^* = \frac{c_i}{\sqrt{2N}}$$

contraints:

$$\sum_j E_{ij} = c_i$$

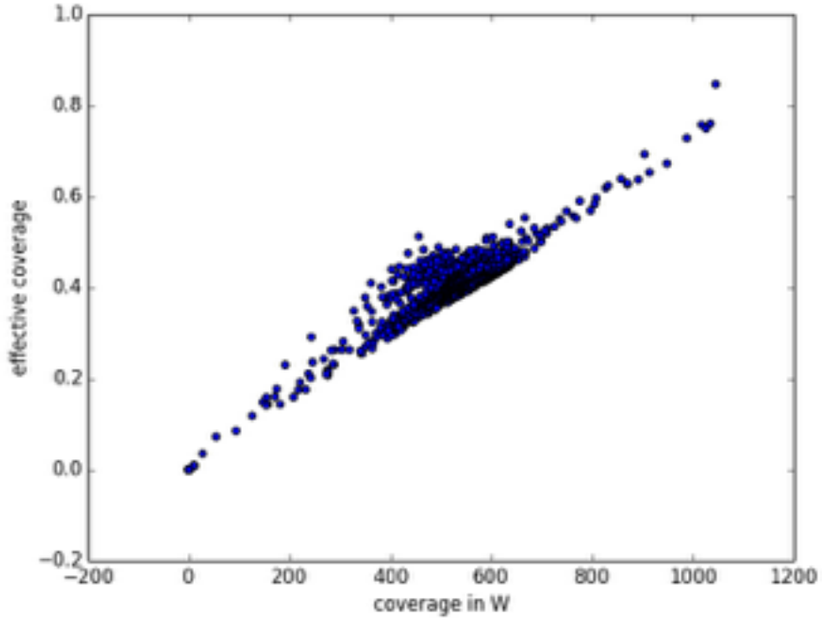
$$\sum_{ij} E_{ij} = 2N$$

genomics distance

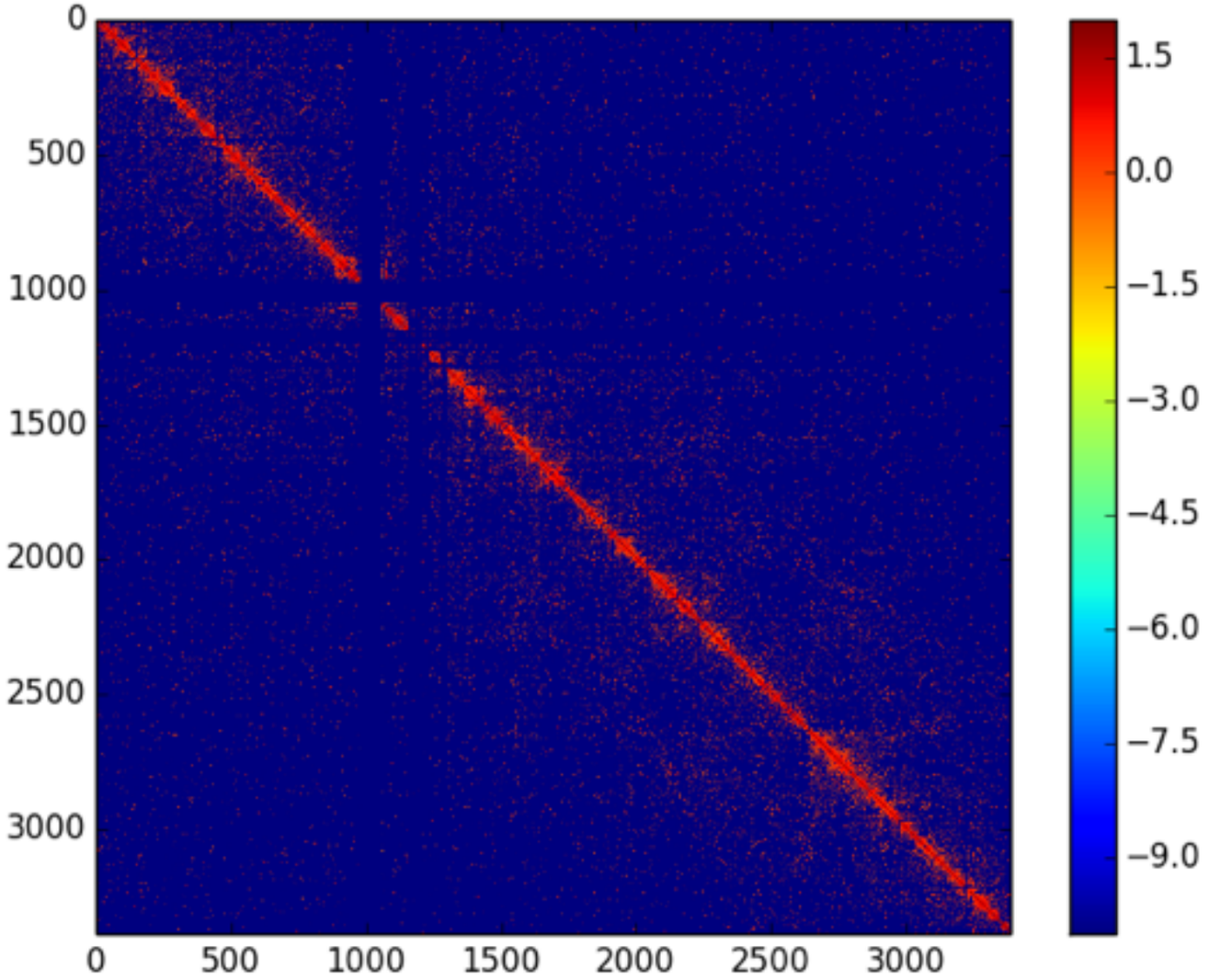
$$E_{ij} = c_i^* c_j^* f(|i-j|)$$

solve c_i^* by iteration

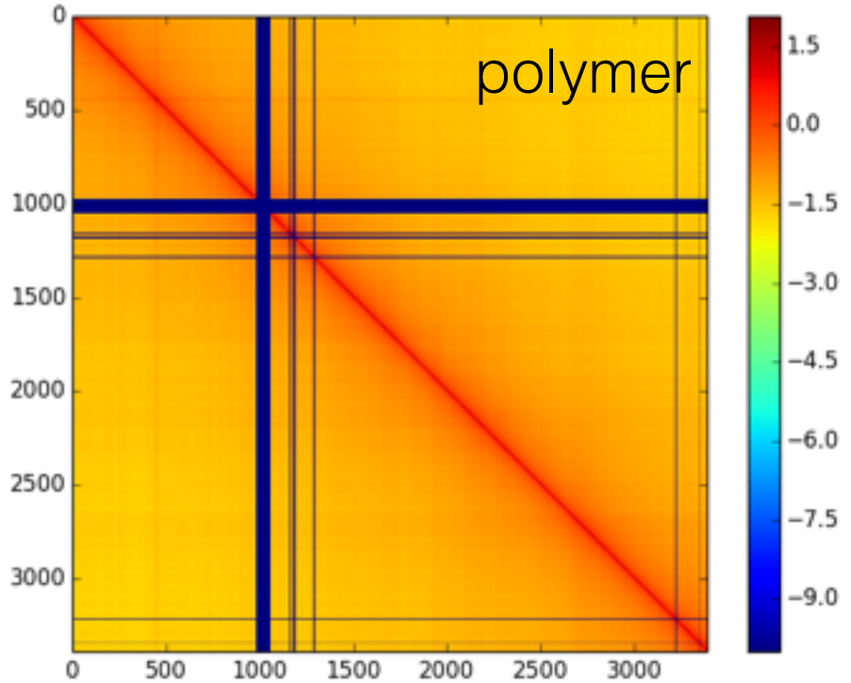
$$Q = \frac{1}{2N} \sum_{ij} (W_{ij} - \gamma E_{ij}) \delta_{\sigma_i \sigma_j}$$

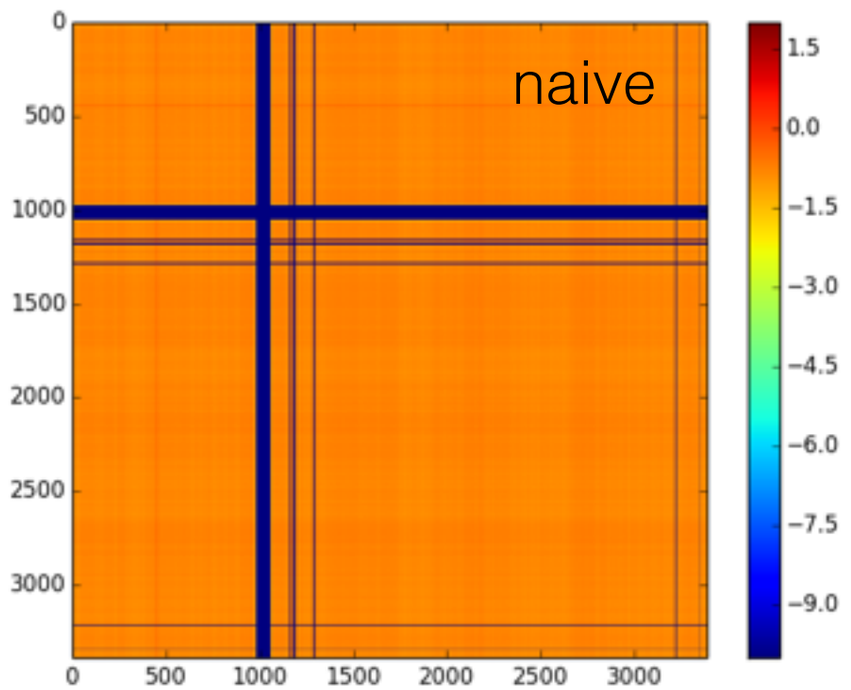


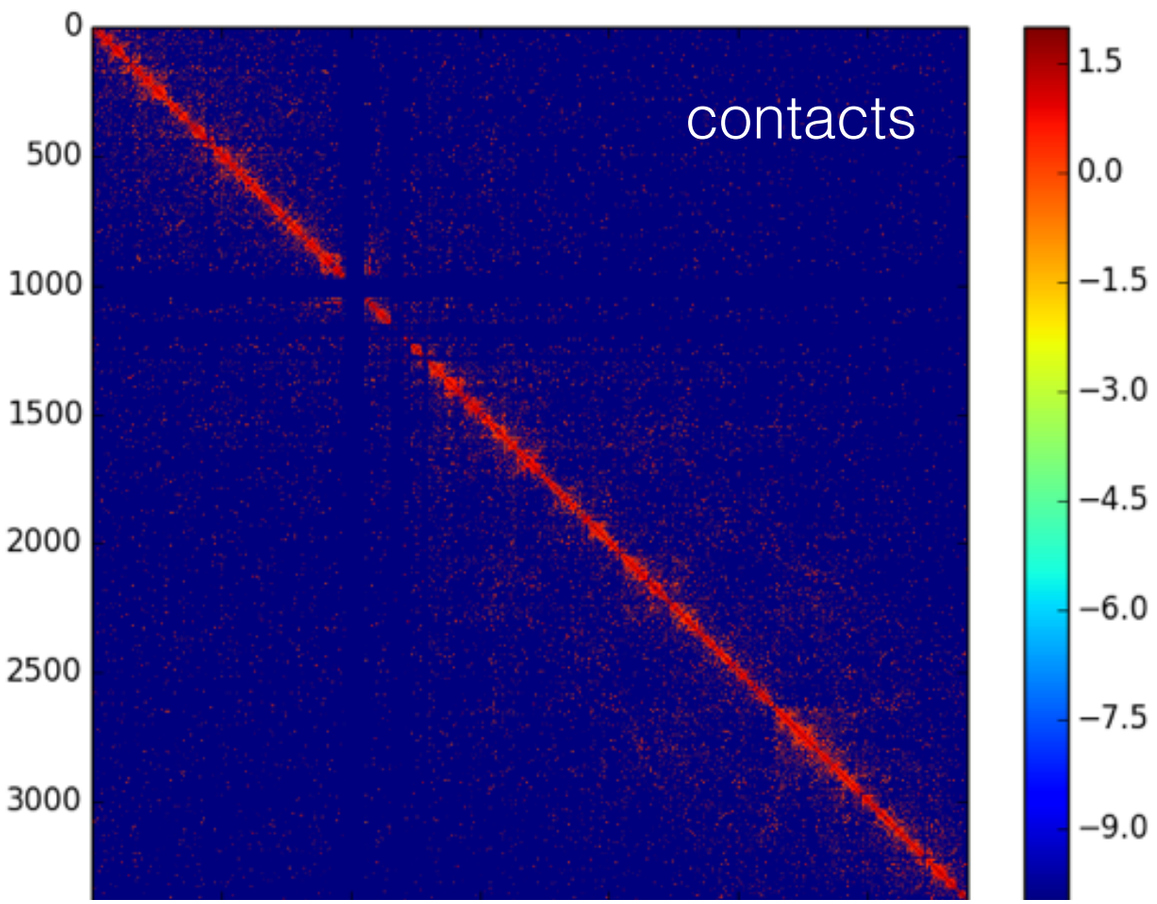
Comparing null models

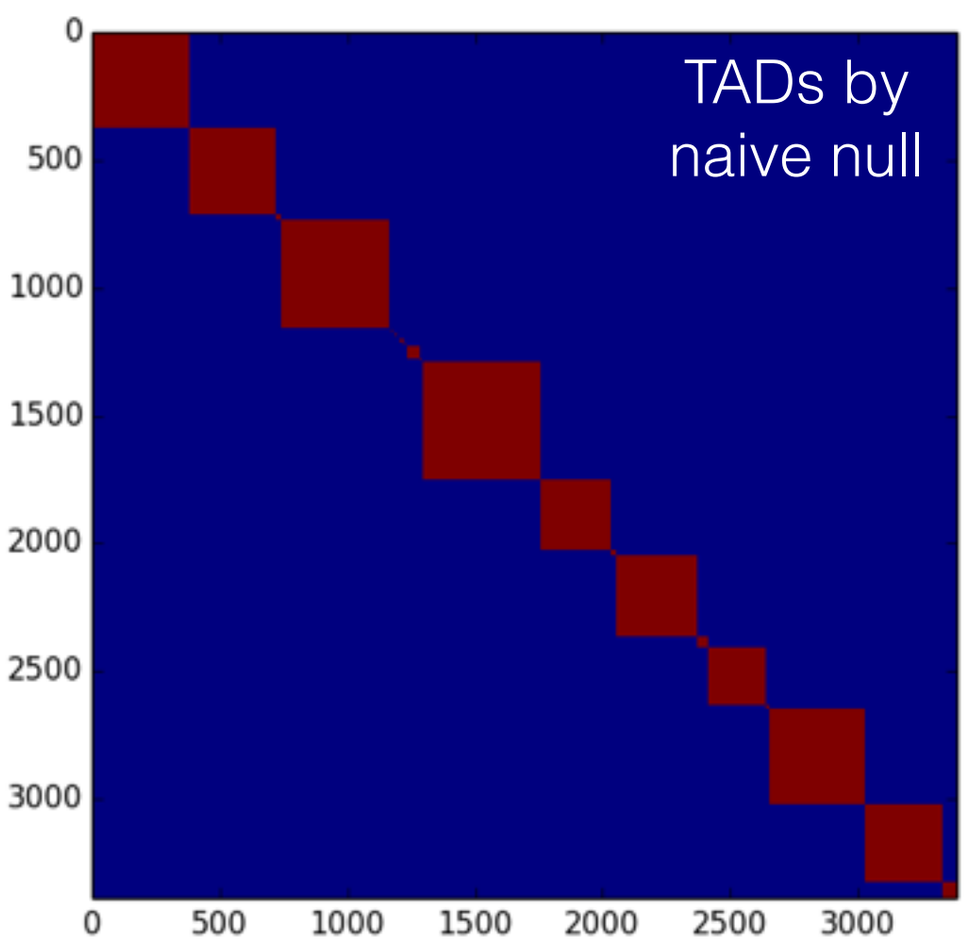


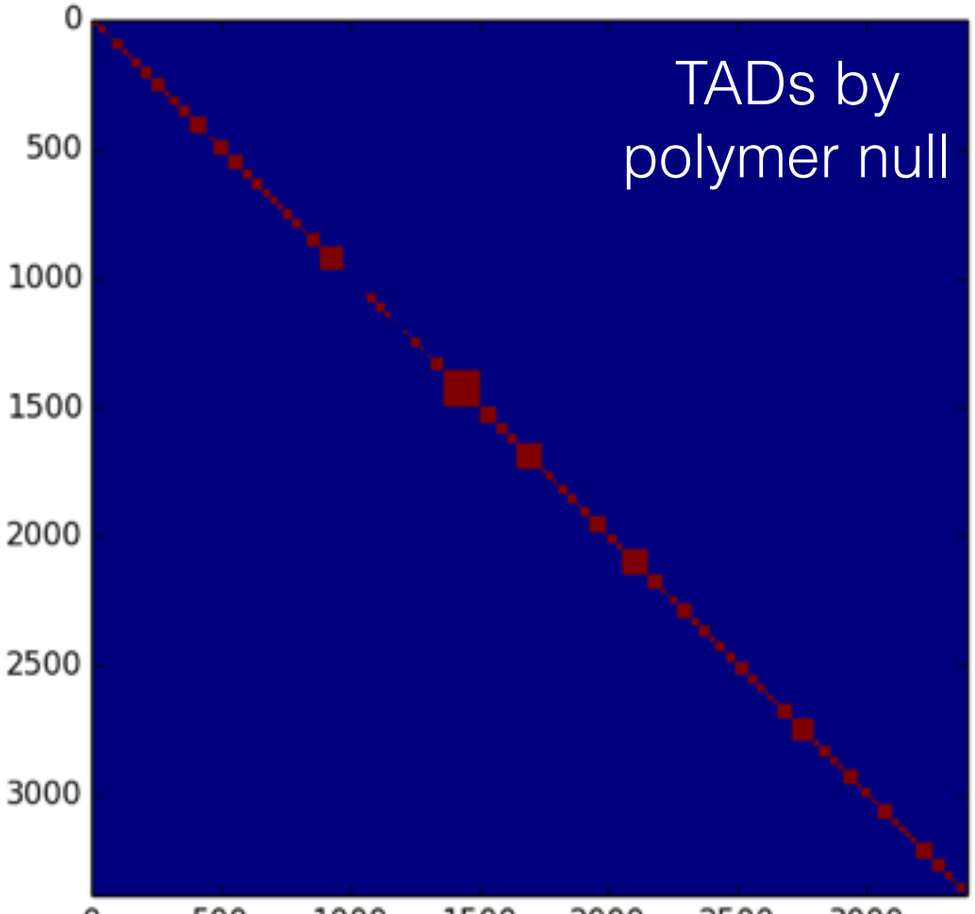
hES, chr 10, 40kb resolution



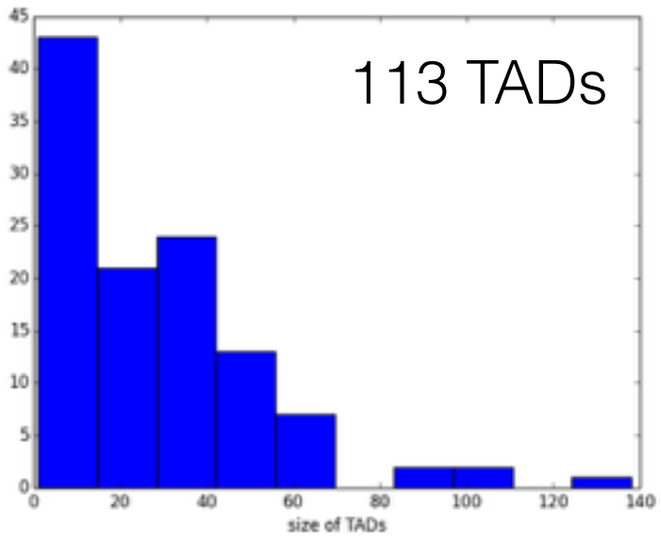


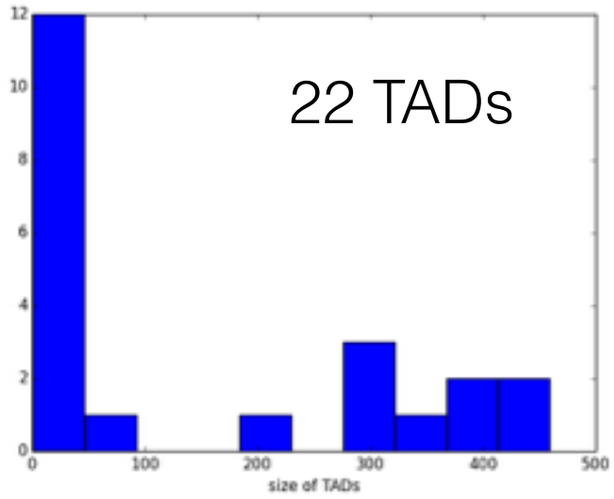






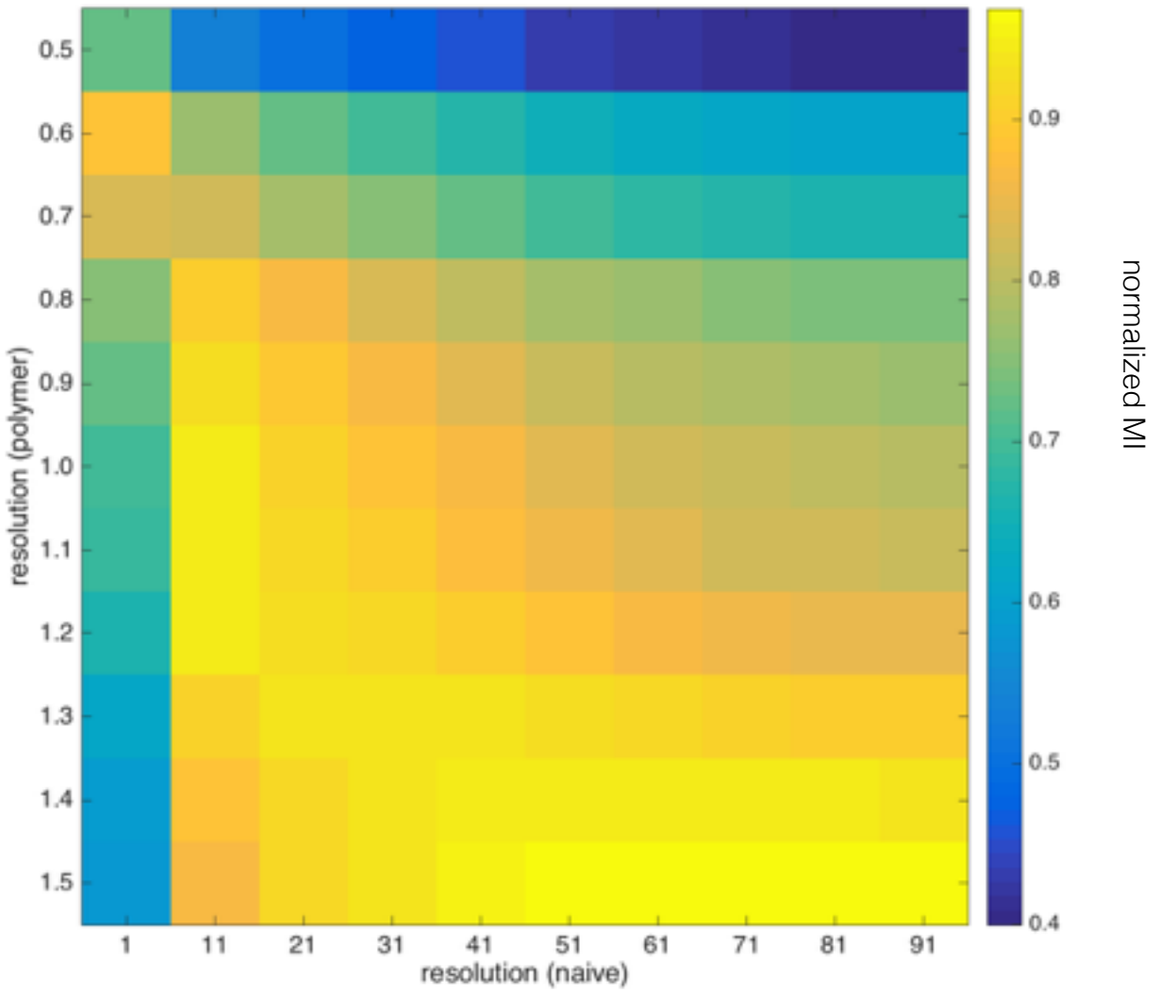
$$Q = \frac{1}{2N}\sum_{ij}(W_{ij} - E_{ij})\delta_{\sigma_i}\delta_{\sigma_j}$$



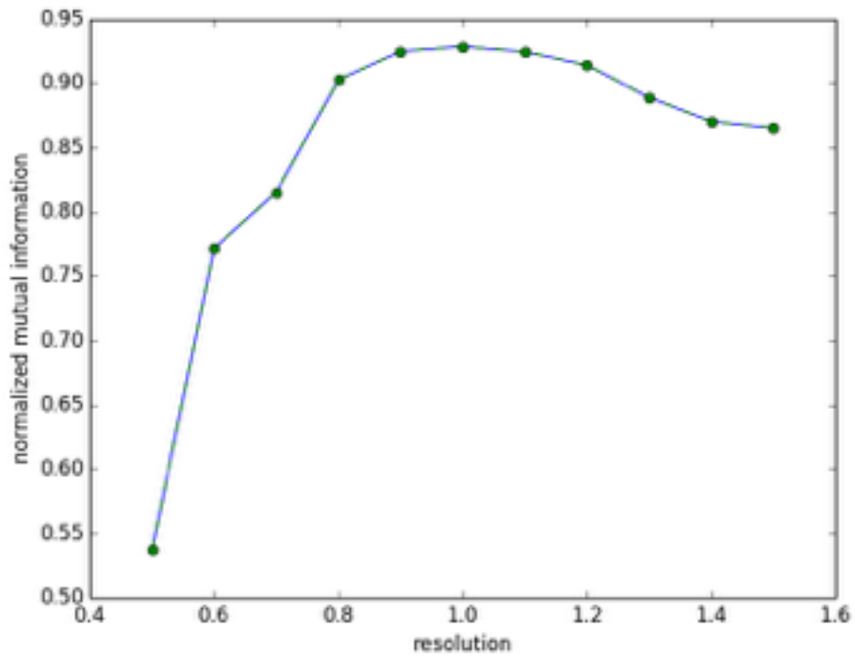


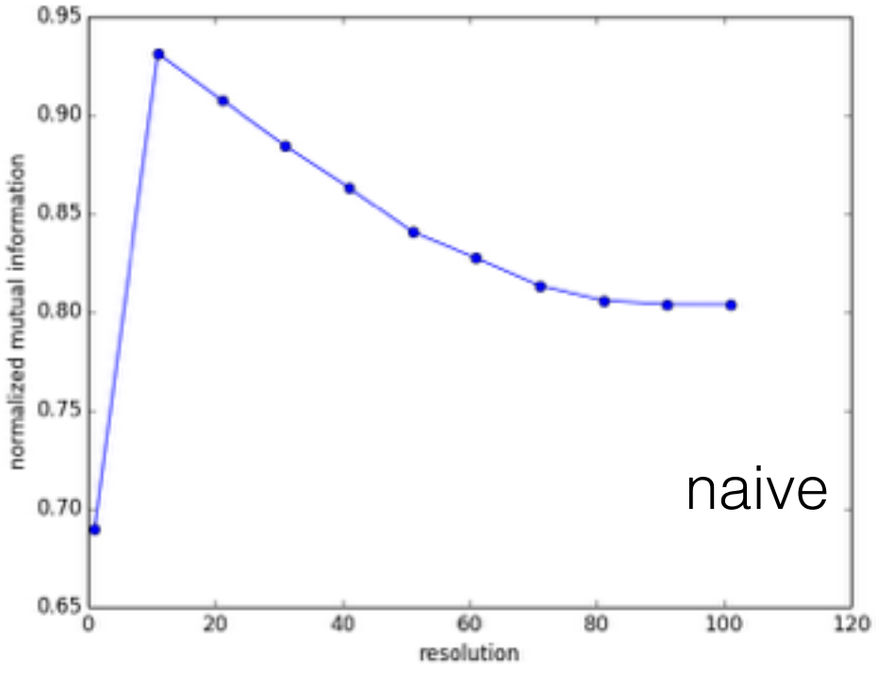
Different resolutions

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

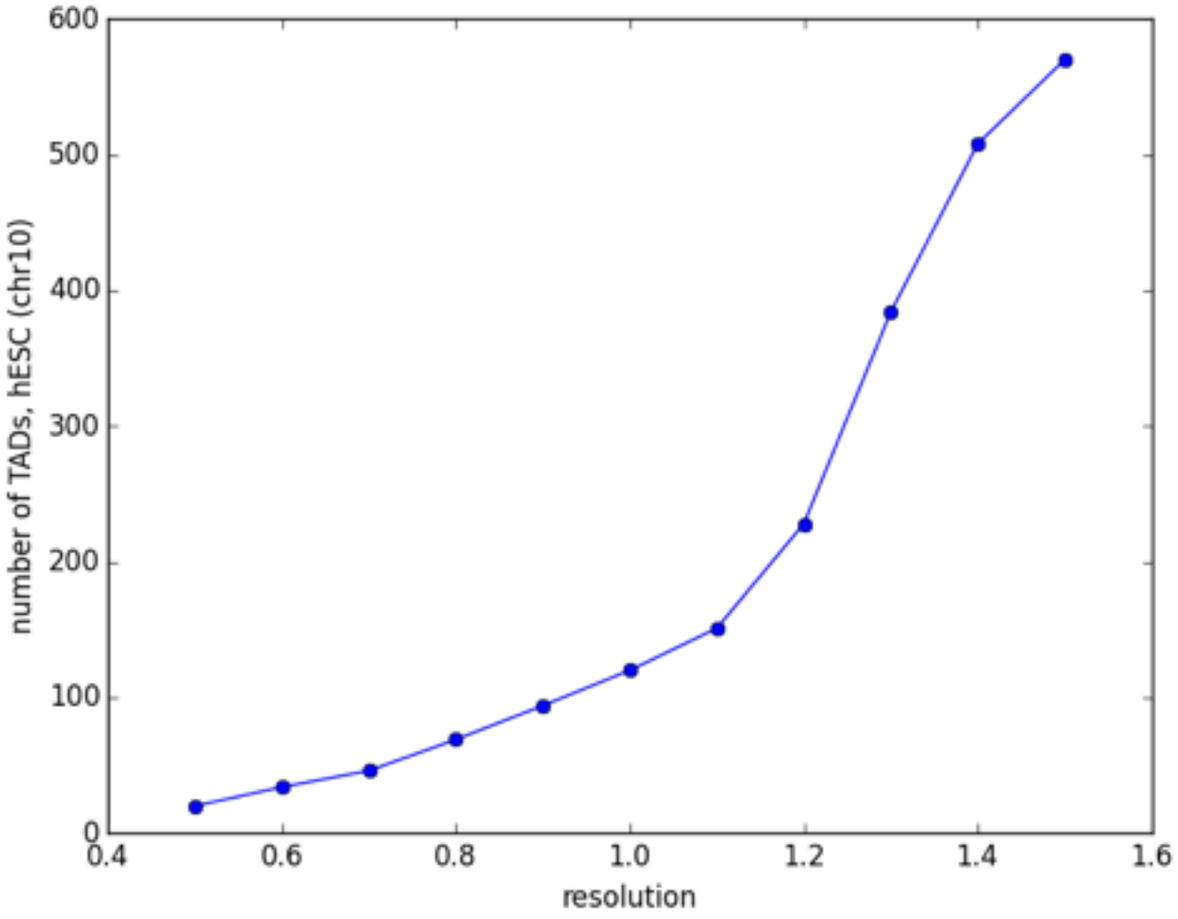


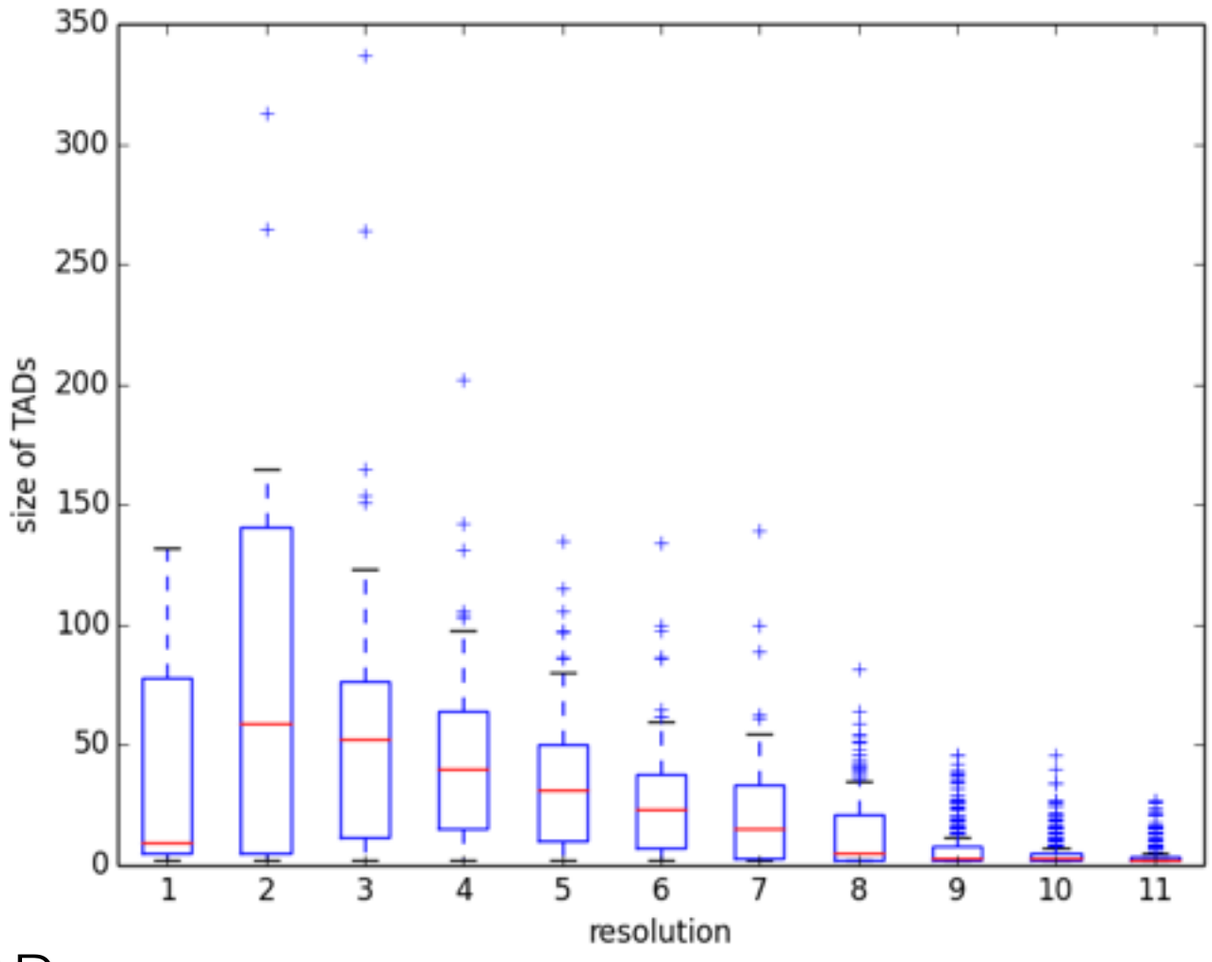
cf. with TADs in Dixon et al.



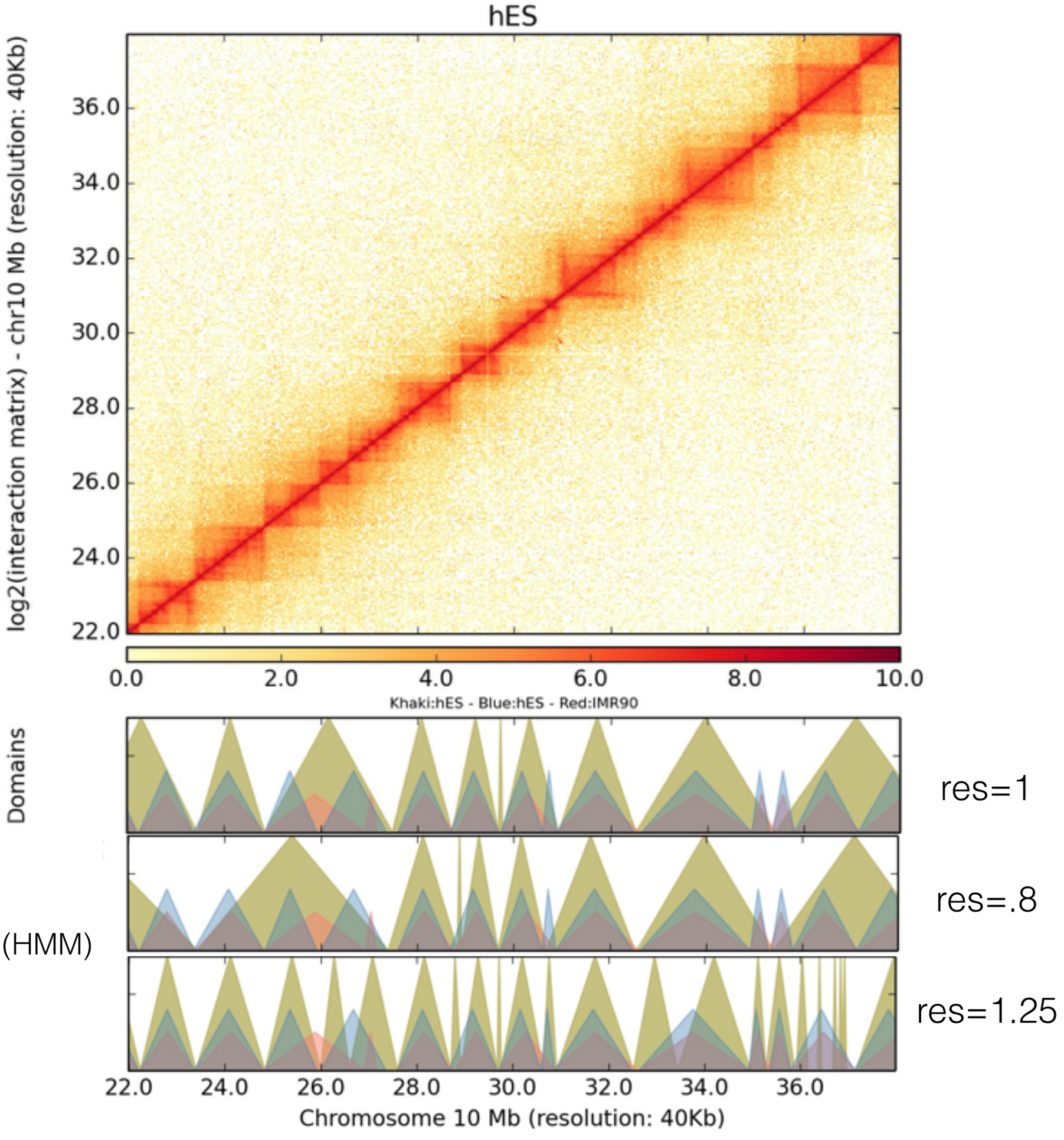


Different resolutions



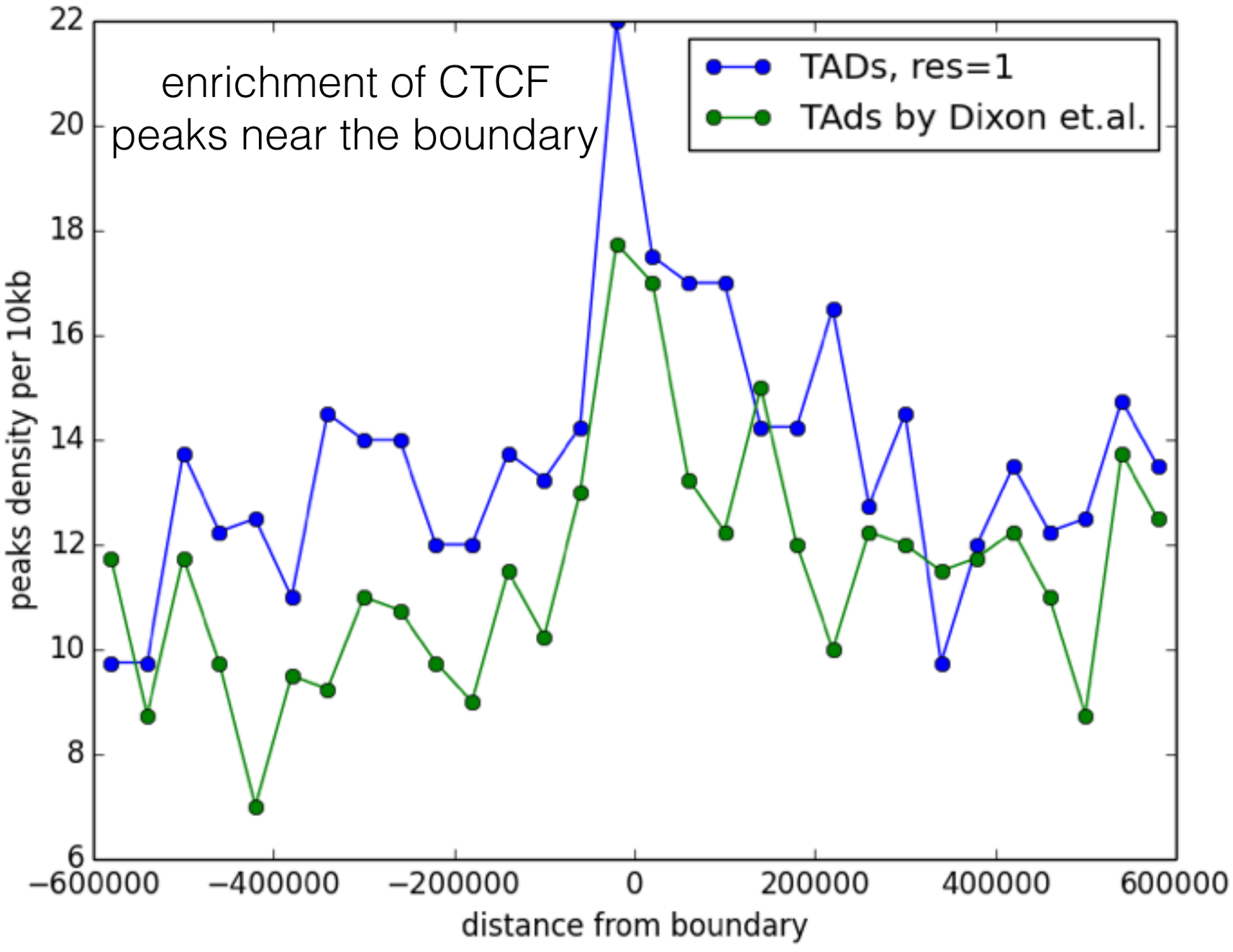


increase resolution gives more but smaller TADs

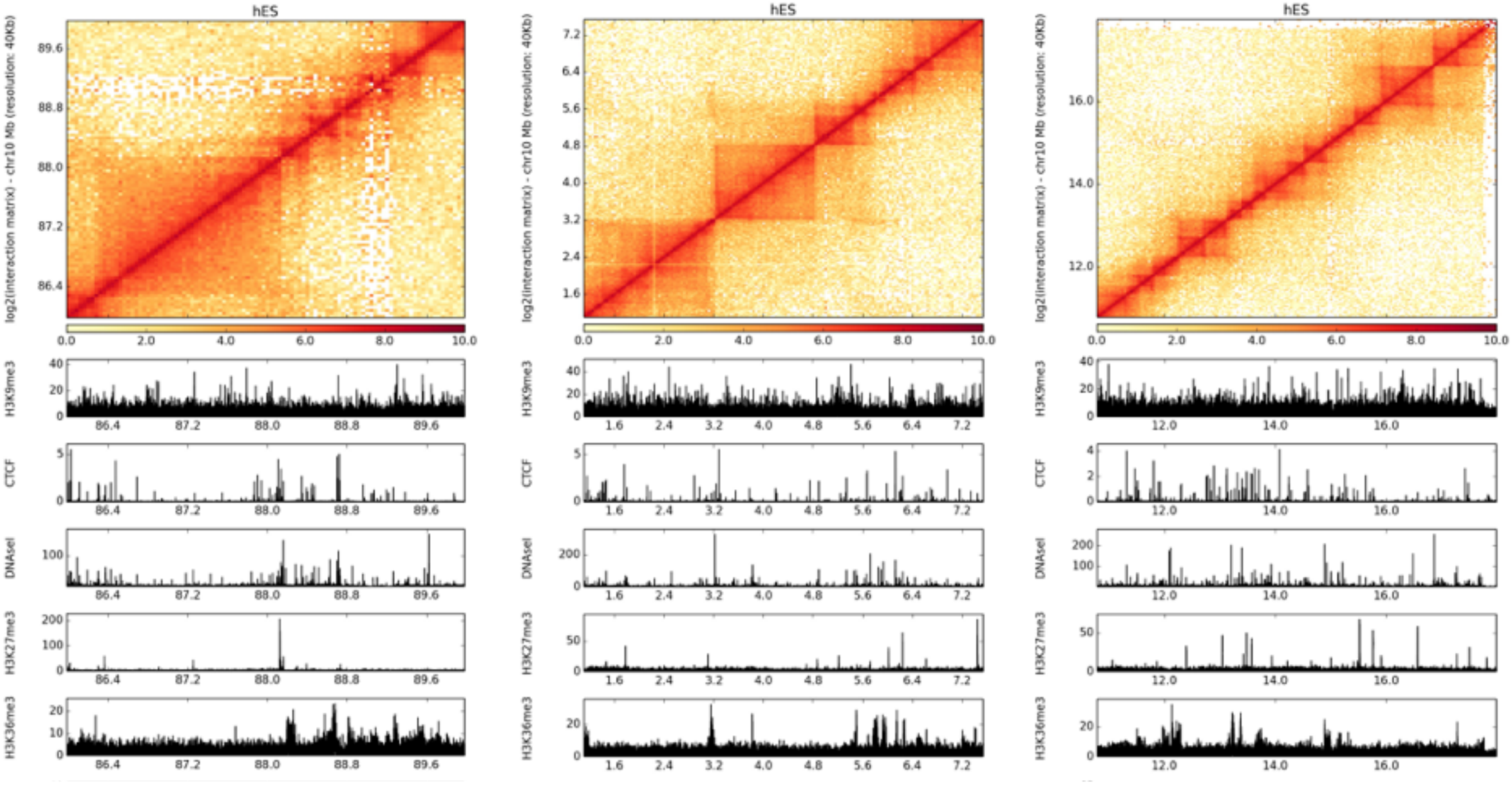


TADs by Dixon et.al. (HMM)
blue: hES
red: IMR90

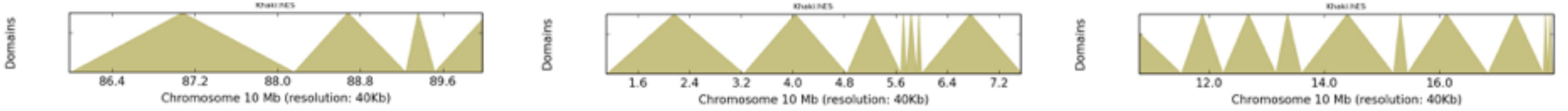
Chromatin features versus TADs



Chromatin features versus TADs



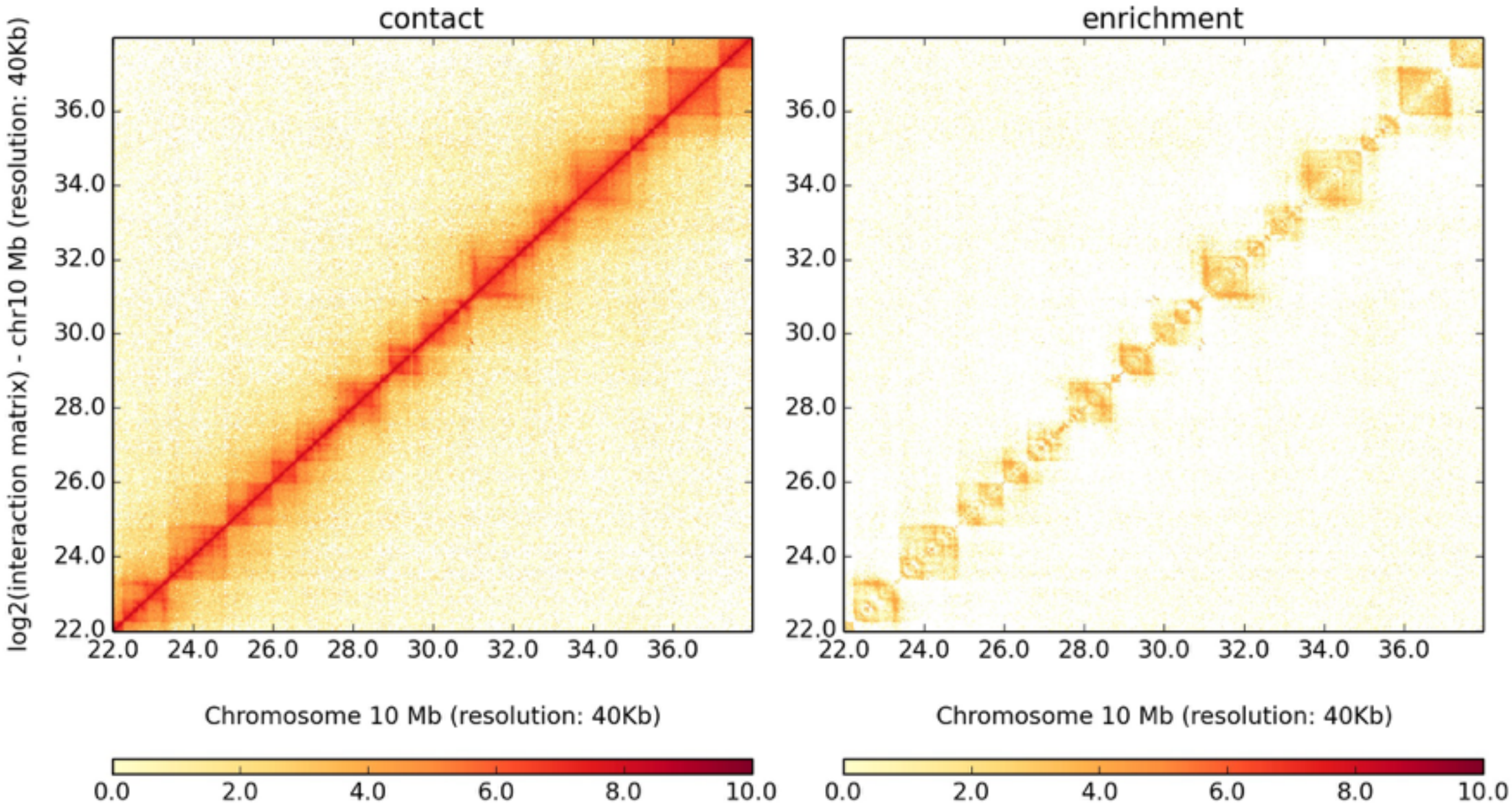
resolution=1 for all three examples



Enrichment of contacts

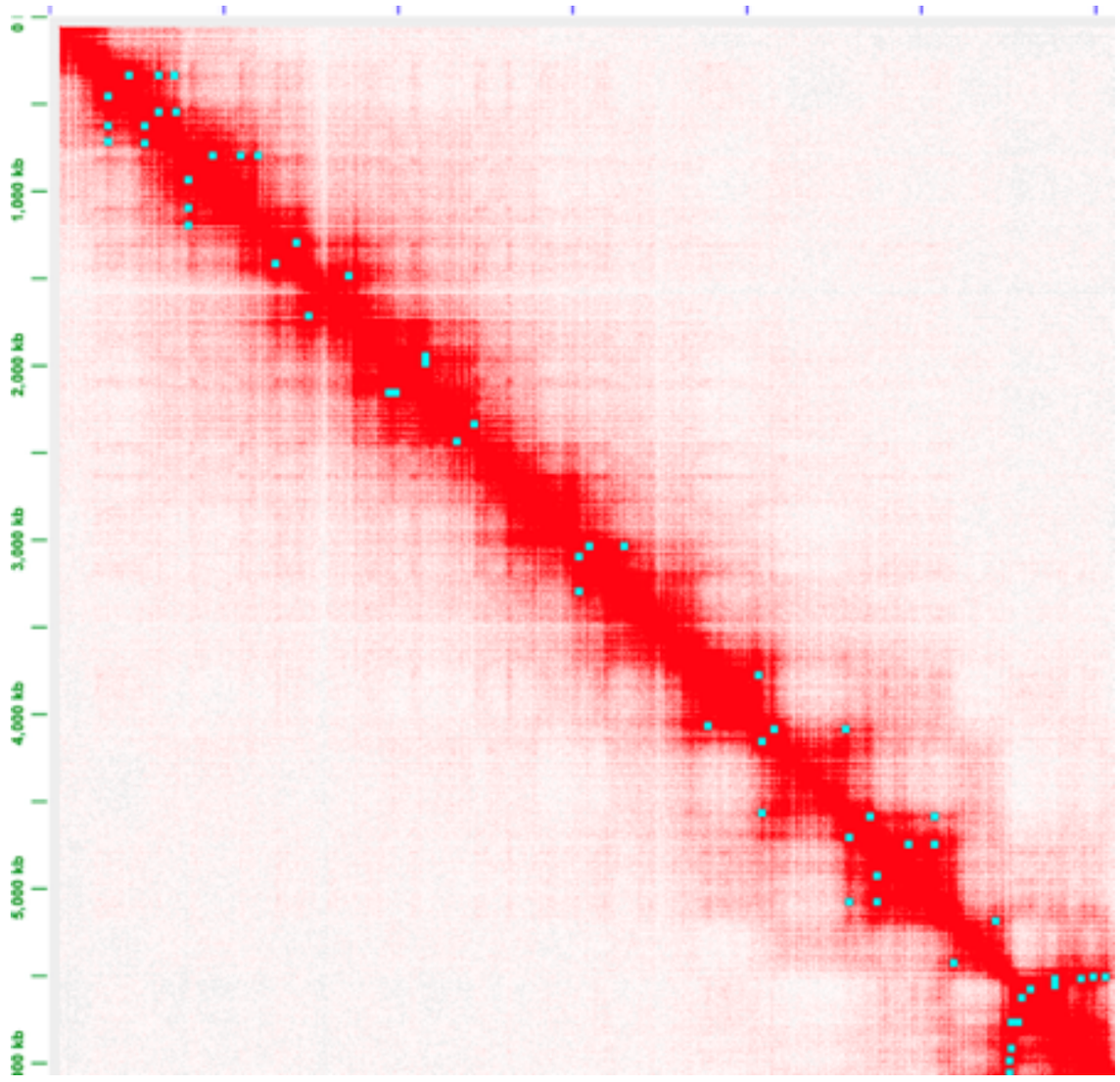
hES cell

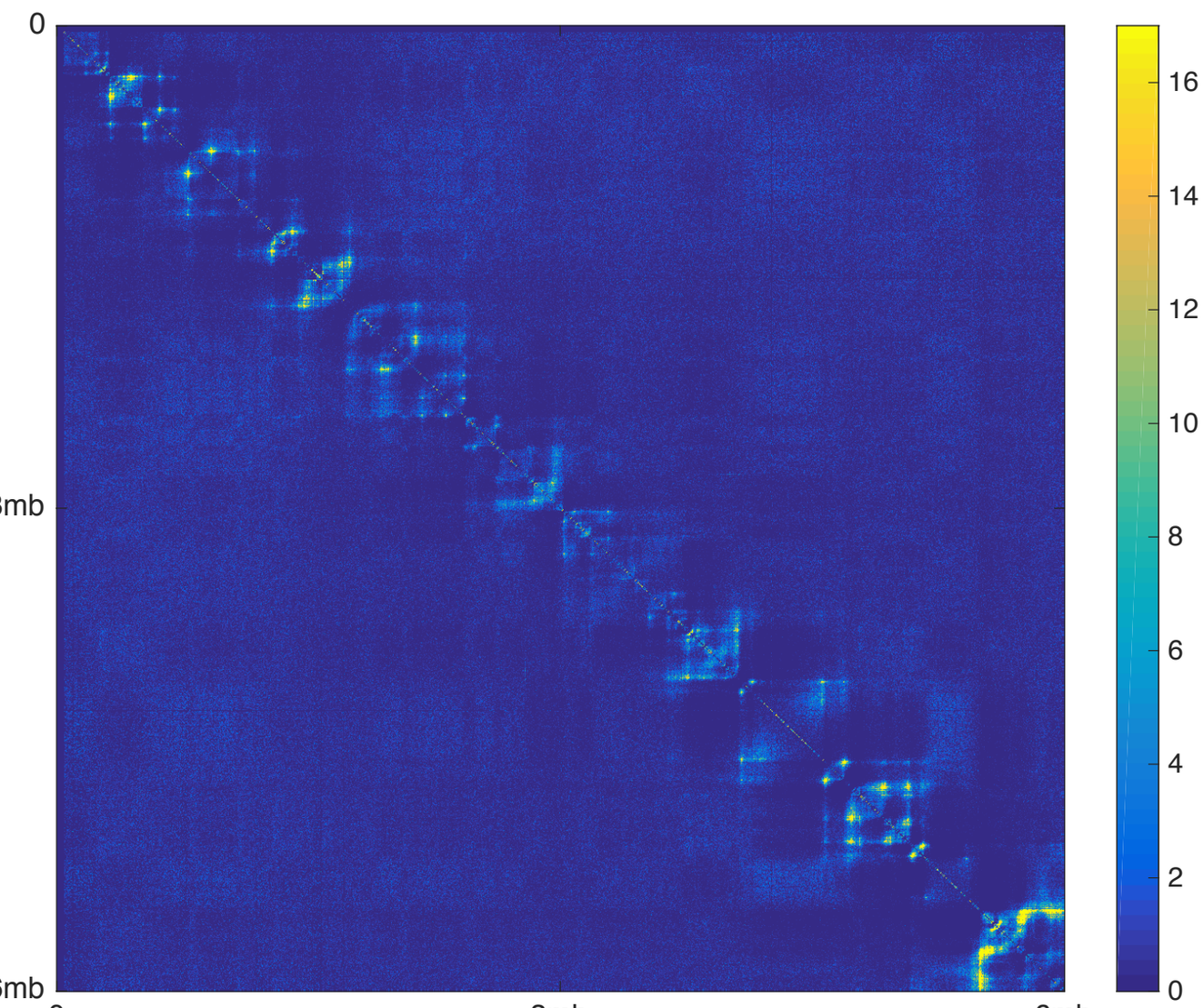
cf. real contacts vs. null
assume Poisson distribution (log P)



Enrichment of contacts

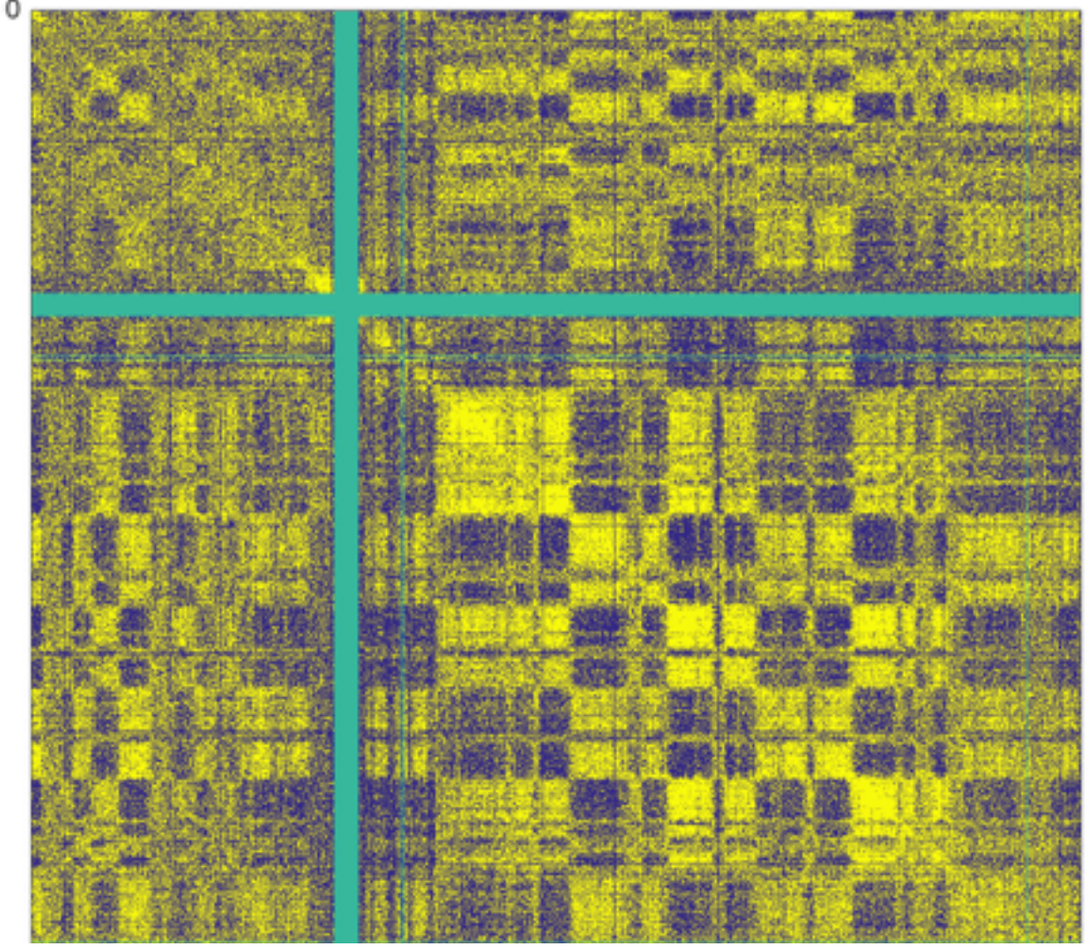
Rao et al. Cell2014, GM12878. bin size=5kb





Compartments

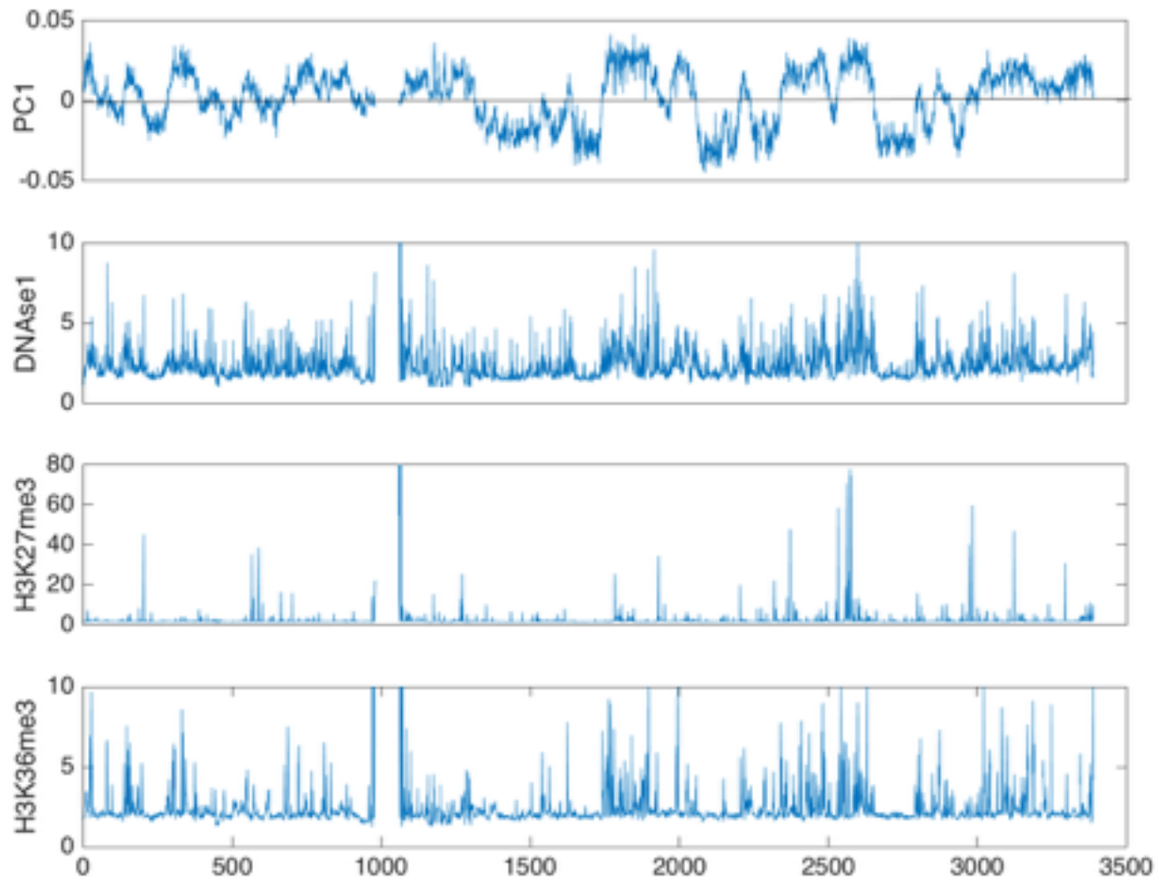
C_ij=corr(observed/expect)



r=0.4982, P=1.00e-205

r=0.3751, P=2.65e-110

r=0.3856, P=0



entire chromosome 10, 40kb binned

Summary and Next steps

- A novel tool to identify TADs

    - Mr TAD Finder (Multi-resolution Topological associated domain) ? Ms TAD Finder (Multi-scale Topological associated domain)?

    - based on global optimization inspired by network modules as oppose to local approaches

    - with a concept of continuous resolution, more general than a hierarchical structure

    - take into account of a background that captures genomic distance

Summary and Next steps

- 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)
        - Gold standards? Signals (e.g. CTCF) near the TAD boundaries
- To investigate the chromatin features of TADs at different resolutions. Different characteristic resolution for different chromatin features.
    - average signal? enrichment? broad peaks?