A network perspective to Hi-C data

Koon-Kiu Yan

Gerstein Lab

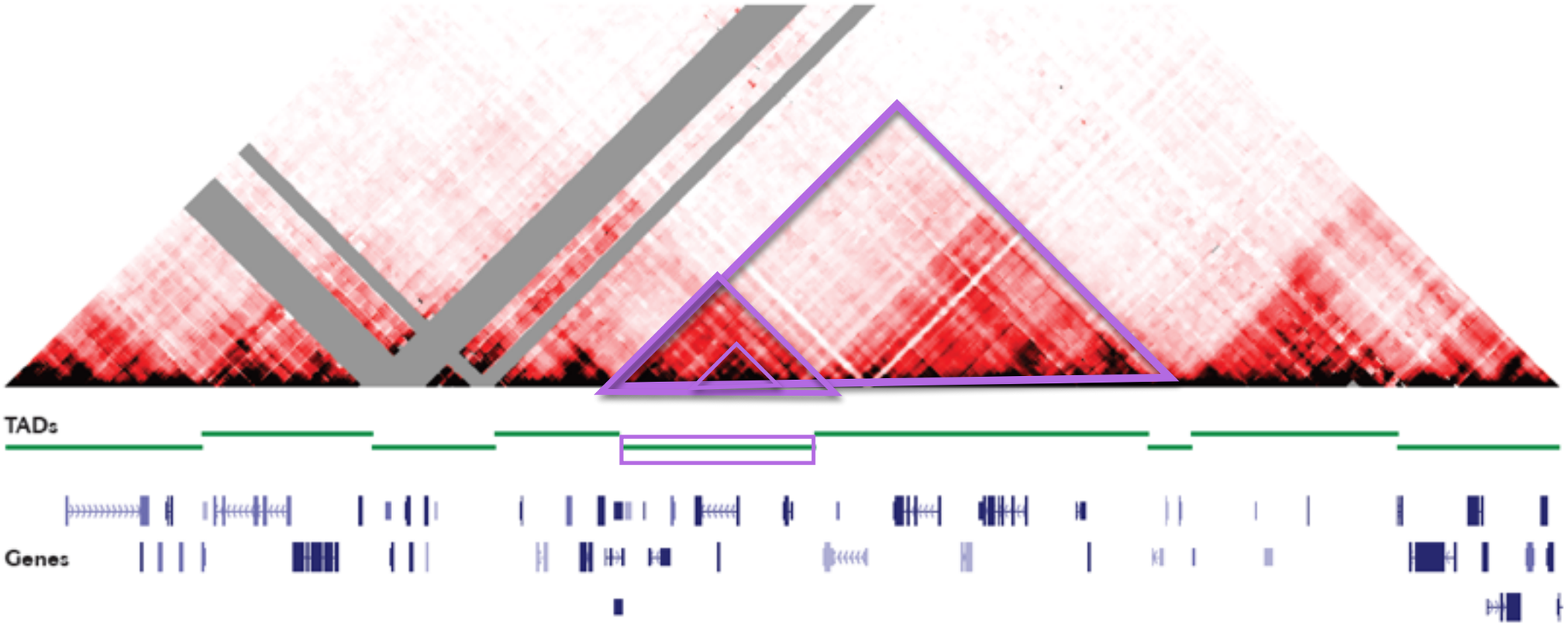
Network provides a system-wide perspective to Hi-C data

- Identifying multi-scale topological domains based on network modularity detection

- A network framework to examine how the spatial organization of genes shapes their expression patterns

- Data used: hES data from Dixon et al.,12 cell lines by Dekker lab

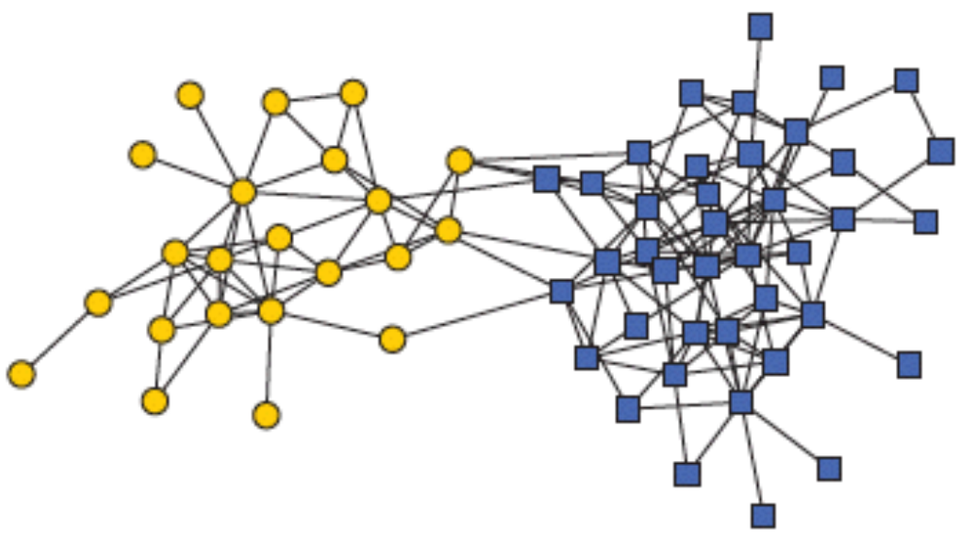
Topologically Associating Domains (TADs)



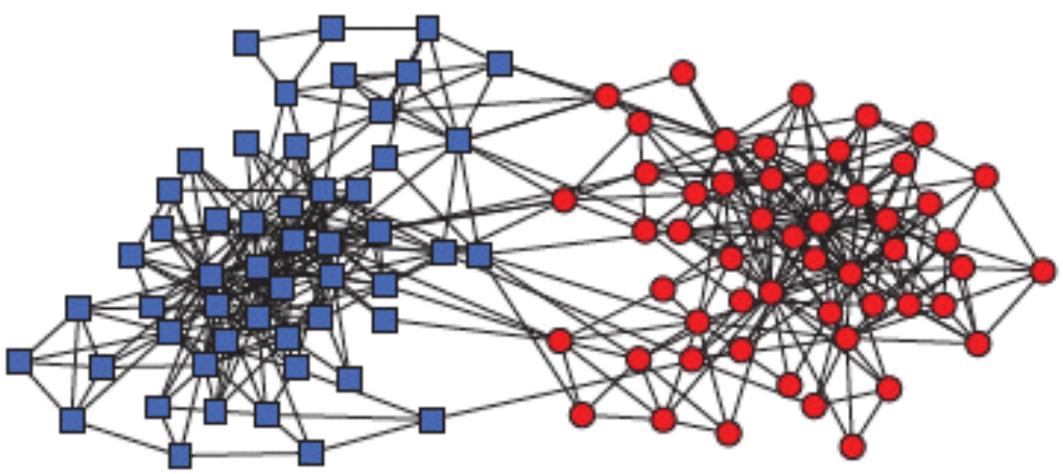
multiple resolutions -> hierarchical organization of genome

Dekker et al. Nat. Rev. Genetics 2013

Network modularity



Dolphin social network



Political books

Newman Phy. Rev. E 2013

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

adjacency matrix

degree of i

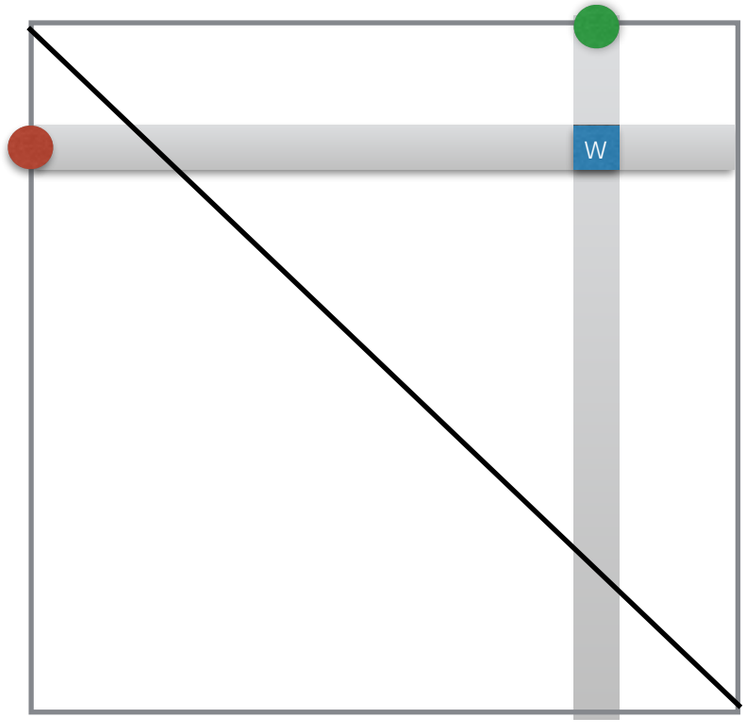
whether or not i, j are in the same module

number of edges

expected number of edges between i and j

Finding TADs based on modularity

Hi-C contact matrix



aRi

aRj

N: the total number of reads

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

expected number of reads between i and j

$$= aR_j * C_i = \frac{aR_j aR_i}{2N}$$

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

Finding TADs in multiple resolutions

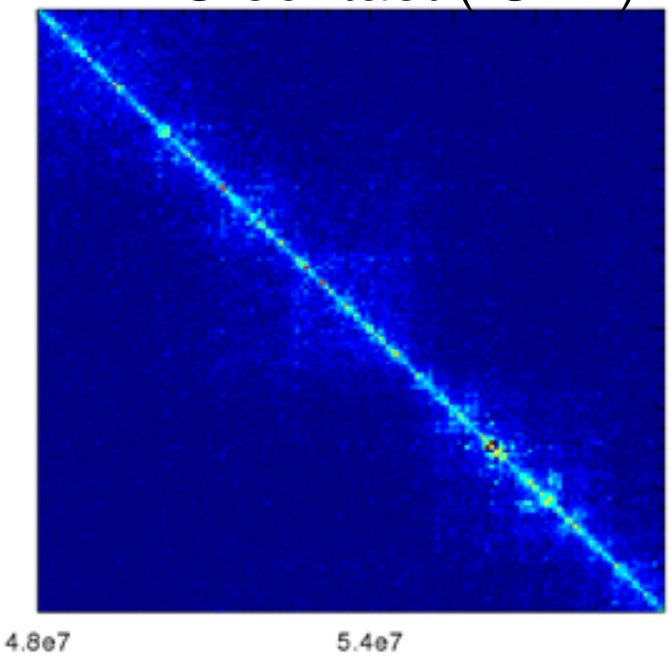
$$Q = \frac{1}{2N}\sum_{ij}(W_{ij} - \gamma\frac{aR_i aR_j}{2N})\delta_{\sigma_i\sigma_j}$$

resolution parameter

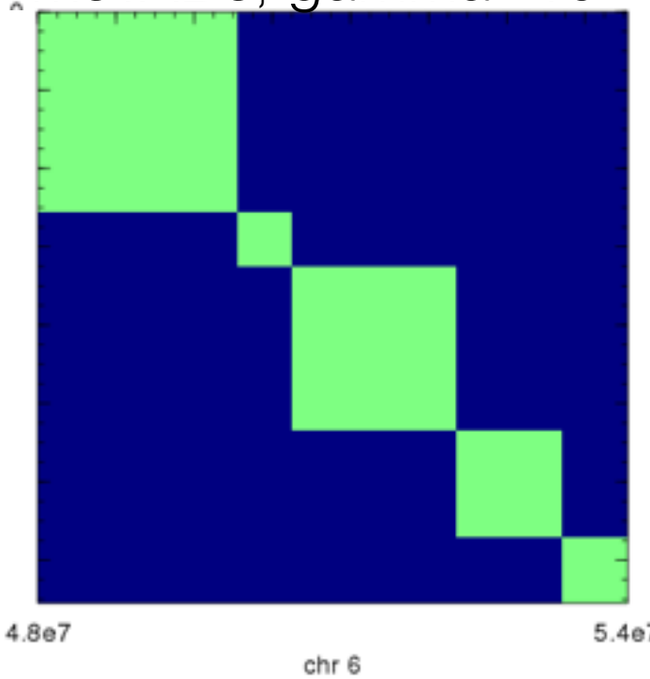
- An increase in gamma results in smaller modules
- An increase in gamma could be interpreted as focusing on the more statistically significant interactions (as compared to the null)
- Input: contact matrix (raw/iced) of the entire genome, or chromosome by chromosome (makes more sense in terms of finding TADs)

Examples (hESC)

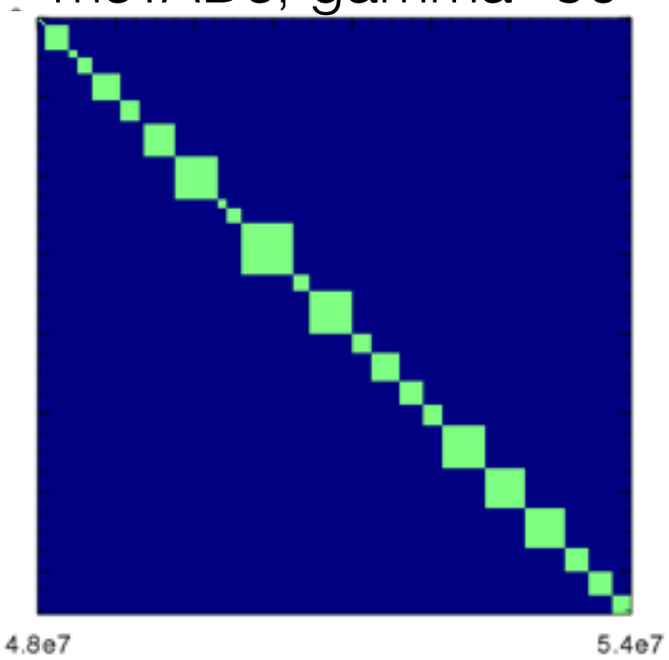
Hi-C contact (ICED)

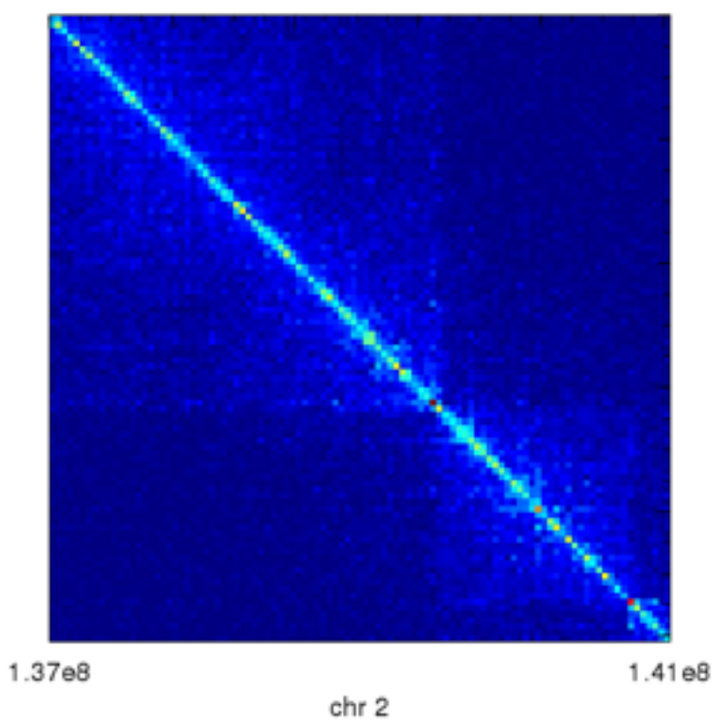


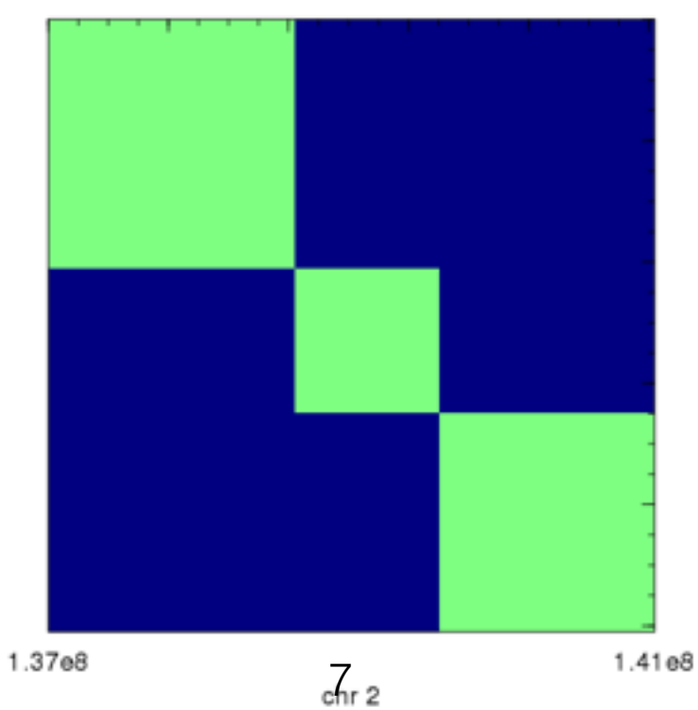
msTADs, gamma=10

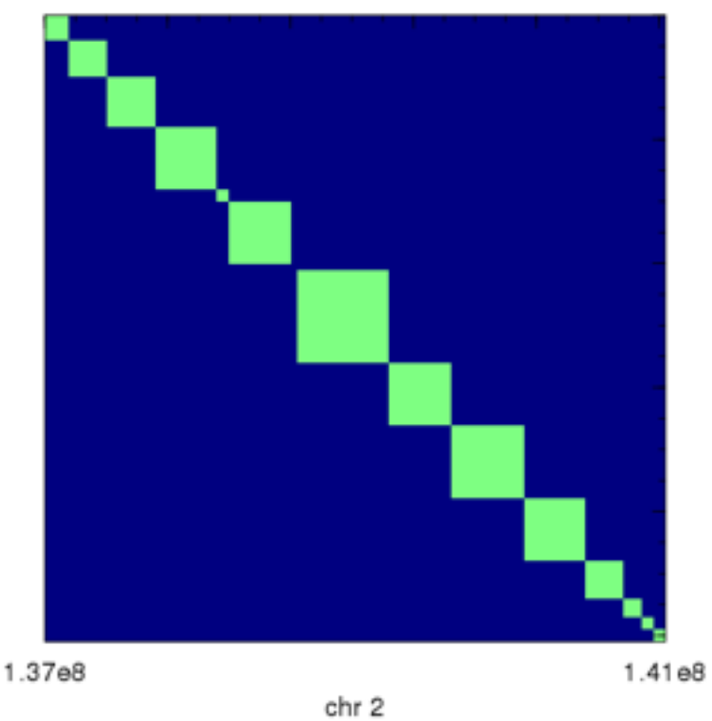


msTADs, gamma=50

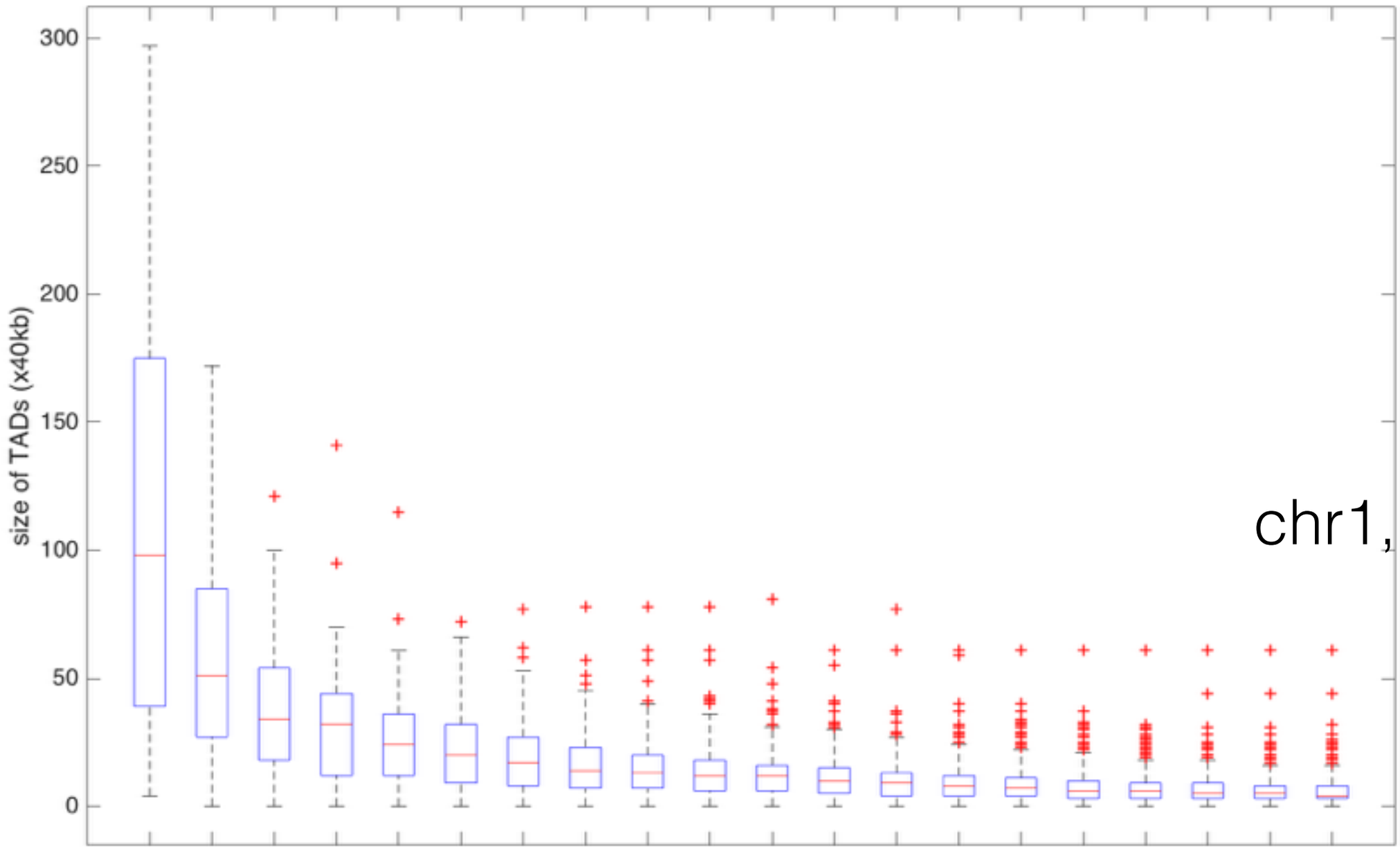








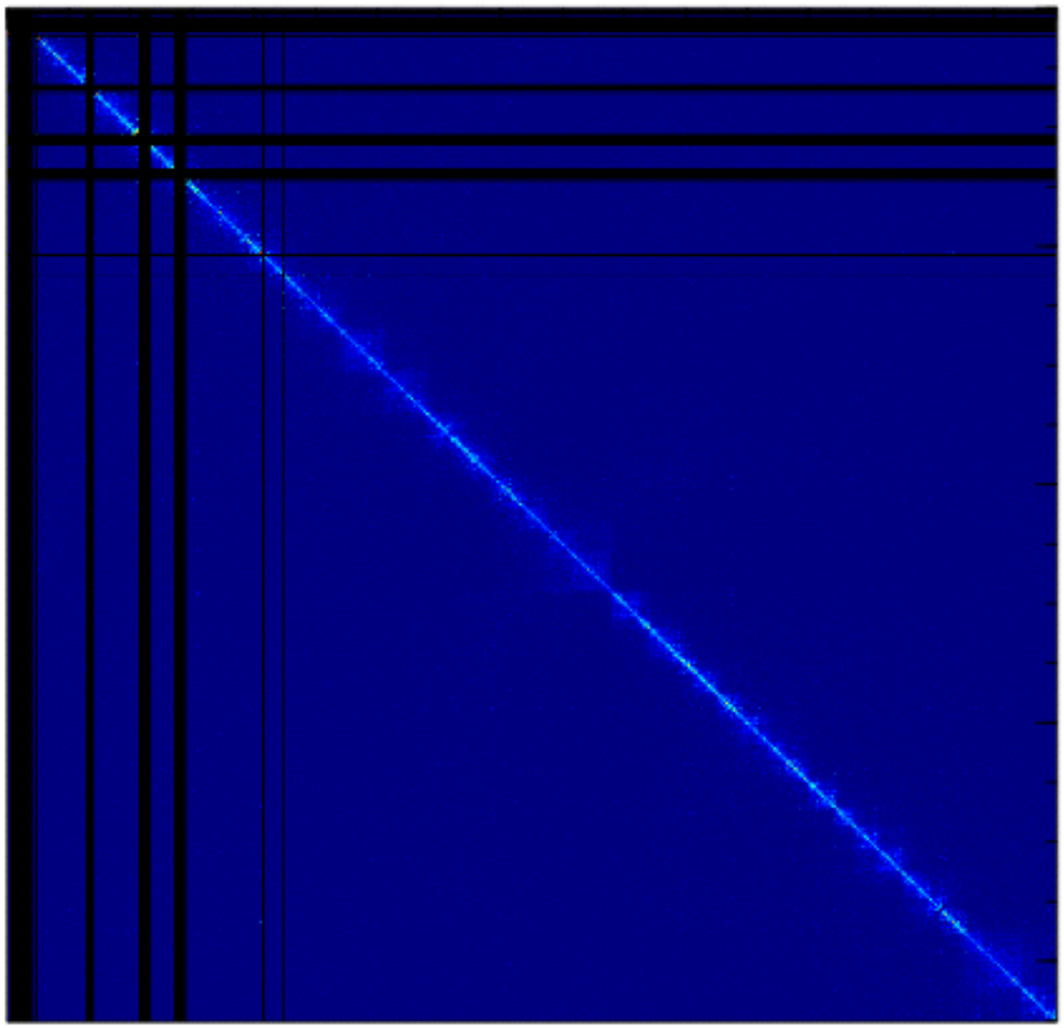
TADs size versus resolution



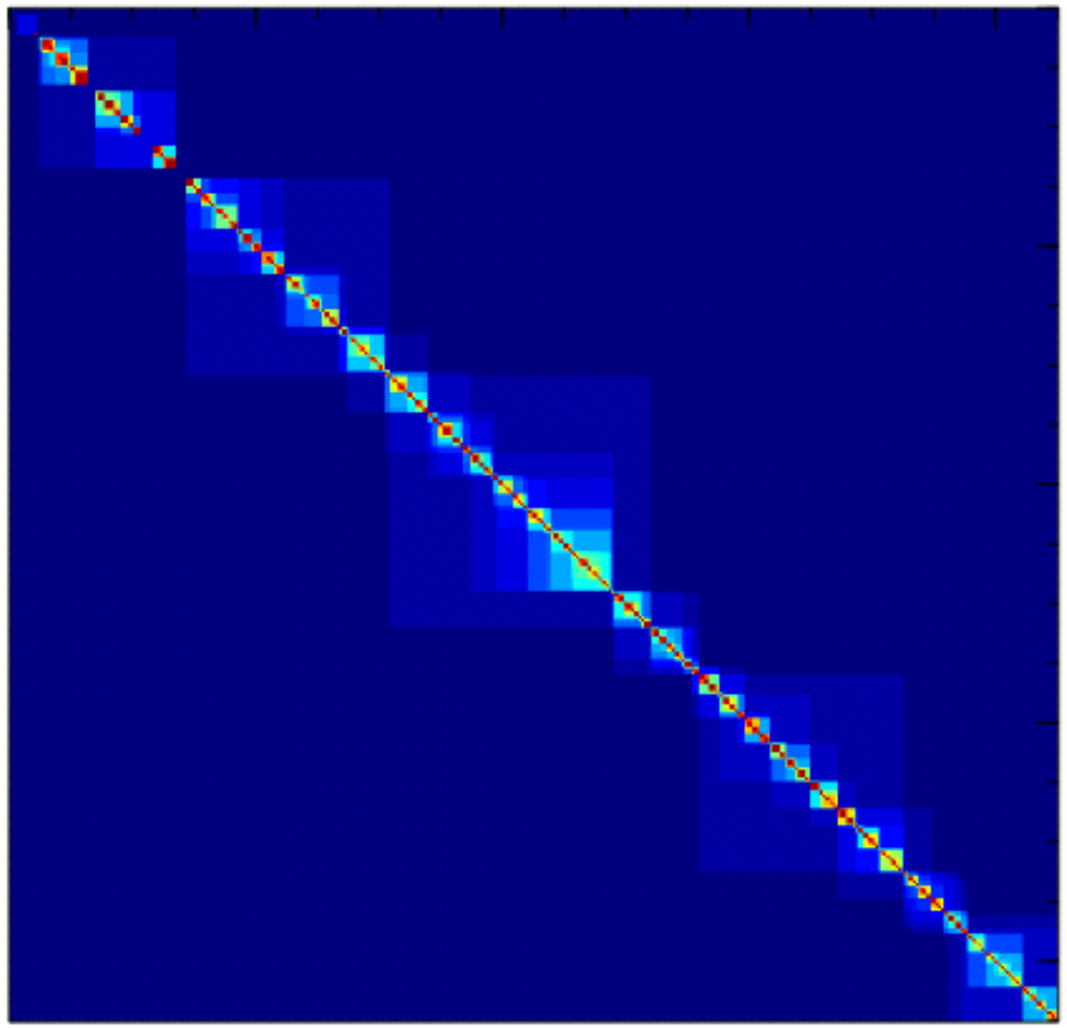
chr1, hESC

Superposing TADs

Hi-C contact (ICED)



msTADs

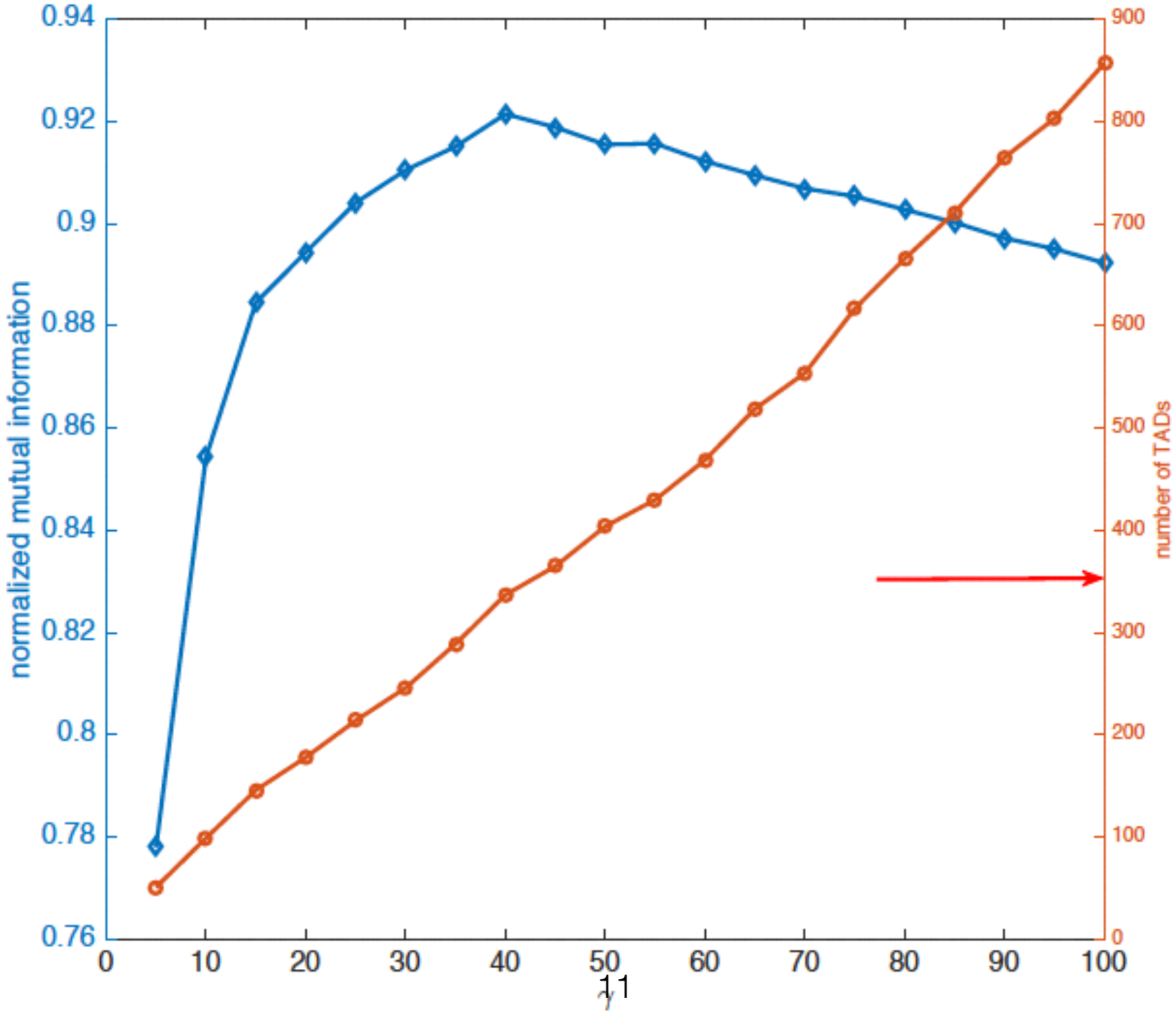


Questions to address

- Is there a characteristic resolution that is the most biologically relevant?

- are there different signatures for different resolutions?

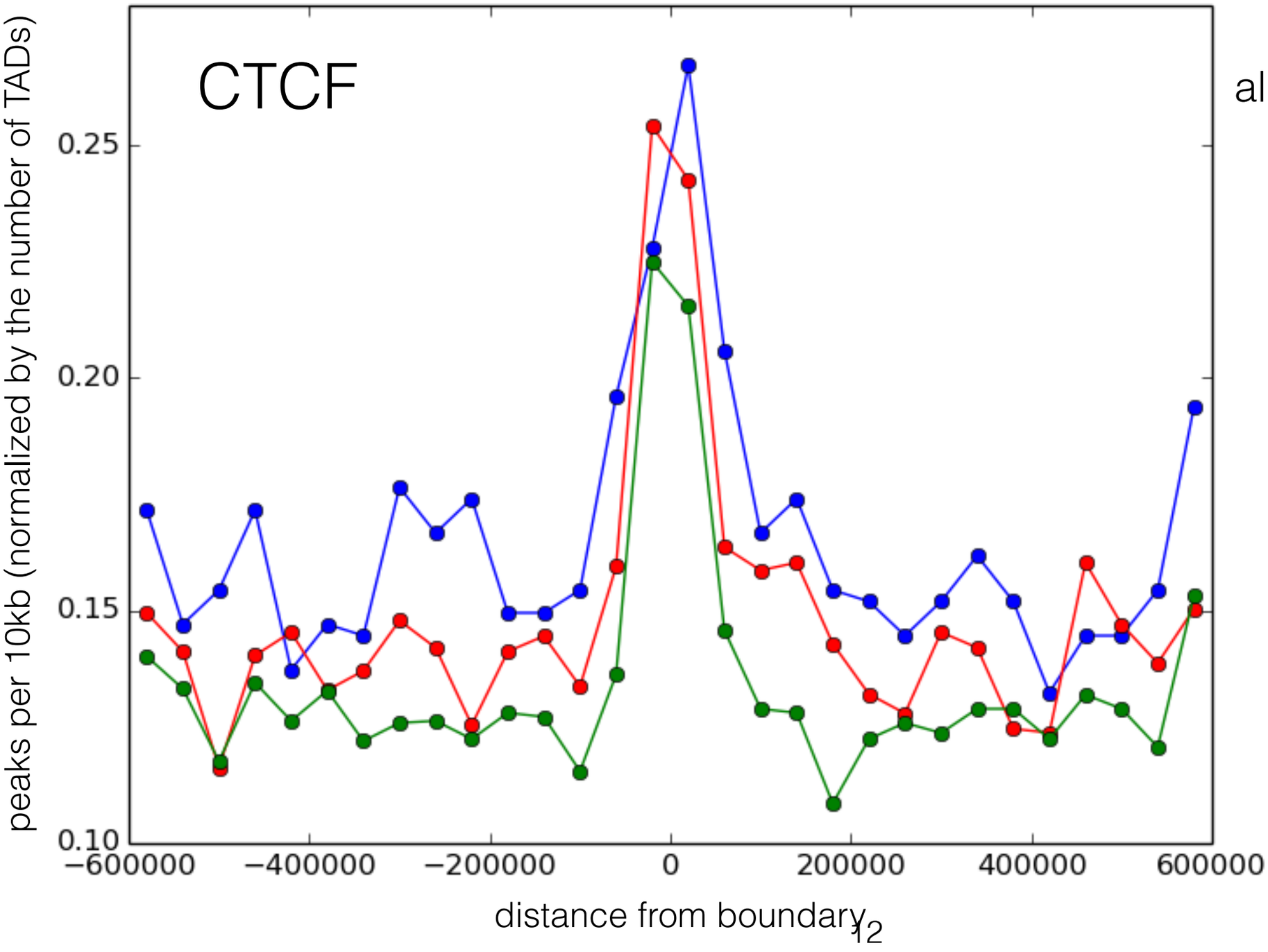
Comparison with HMM method



chr 1 of hESC

TADs based on Dixon et al. 2012

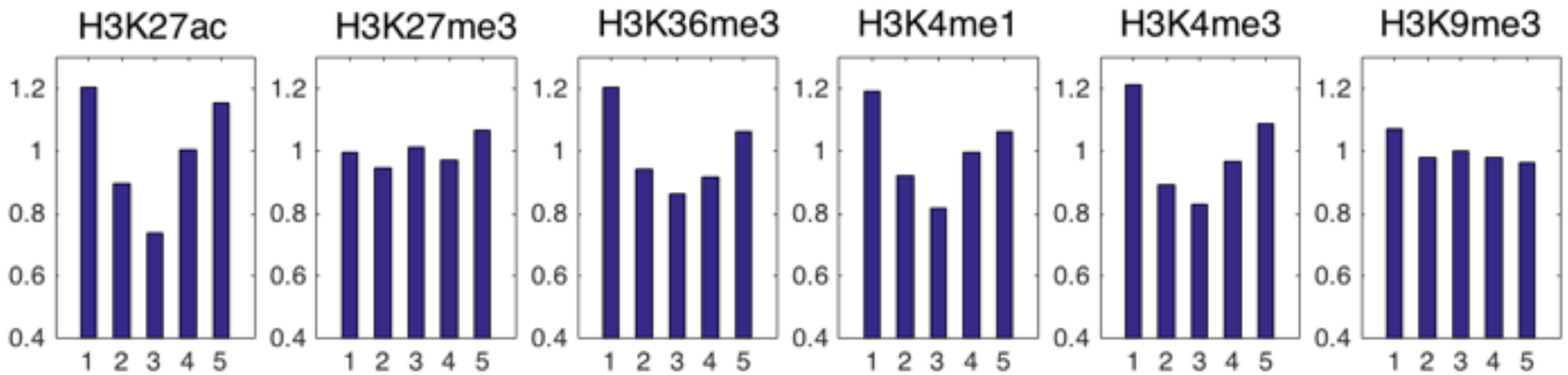
Boundaries between TADs



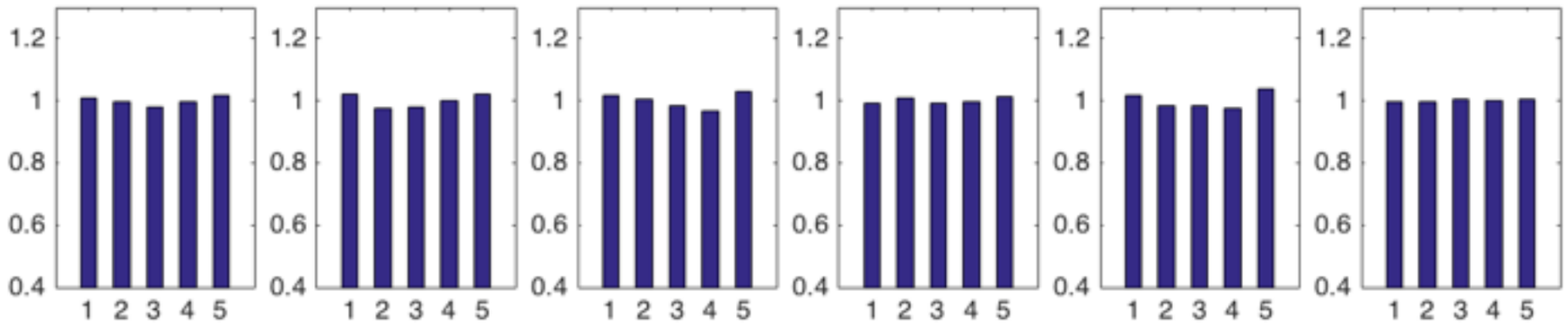
all TAD boundaries in chr1

gamma	num. of boundaries
10	102
30	301
50	578

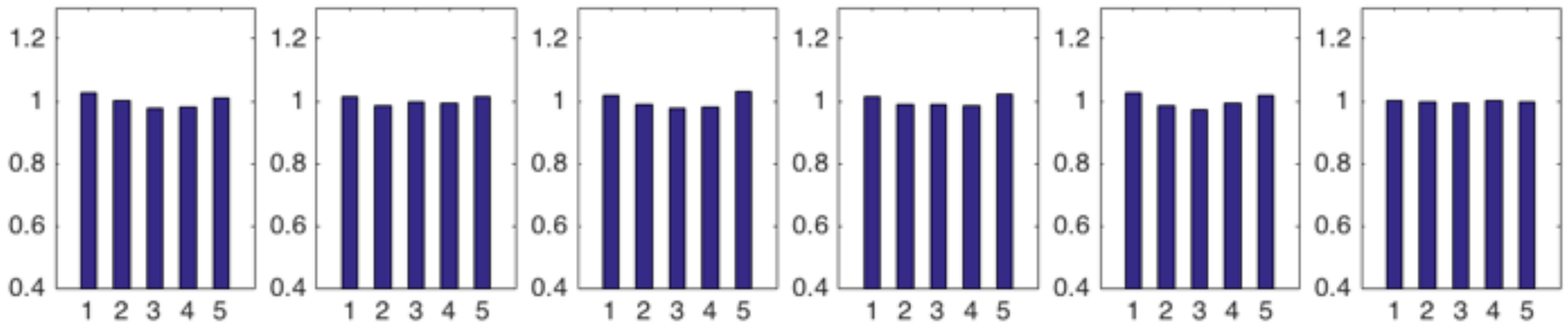
Chromatin signatures for different resolutions



res=10



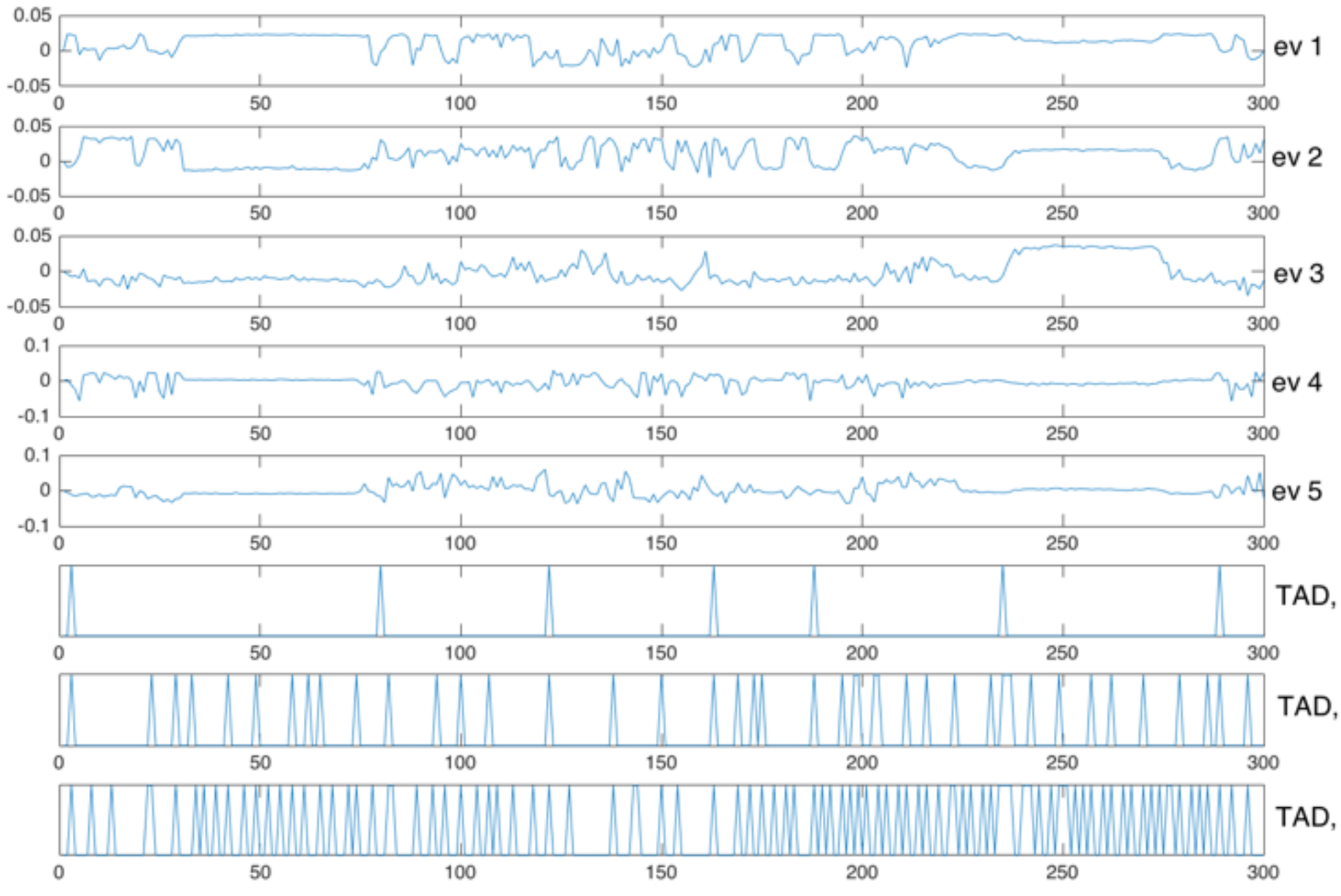
res=40



res=70

Chromatin signatures for different resolutions

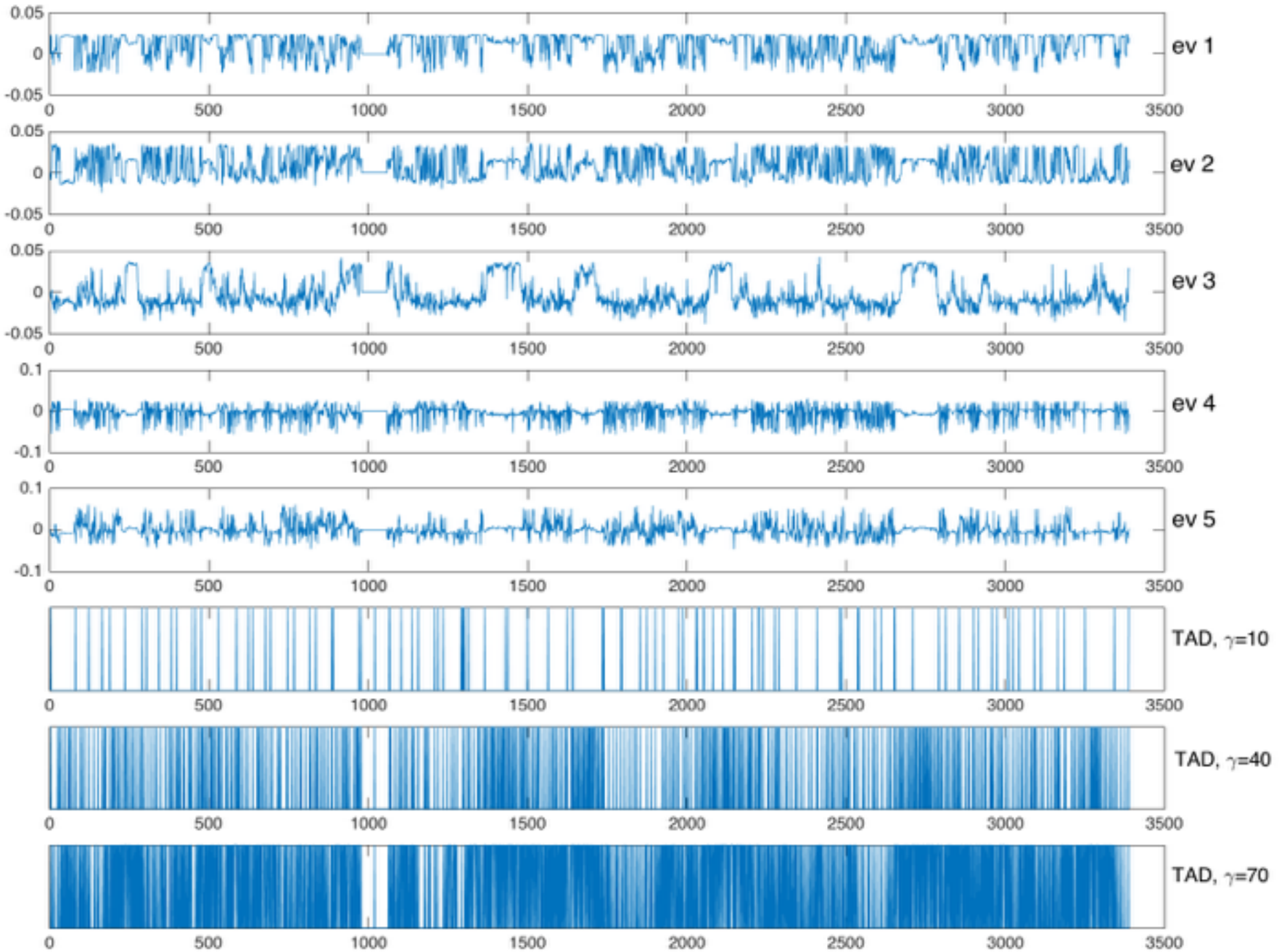
H3K27ac
H3K27me3
H3K36me3
H3K4me1
H3K4me3
H3K9me3



chr. 10, bin size: 40kb

Chromatin signatures for different resolutions

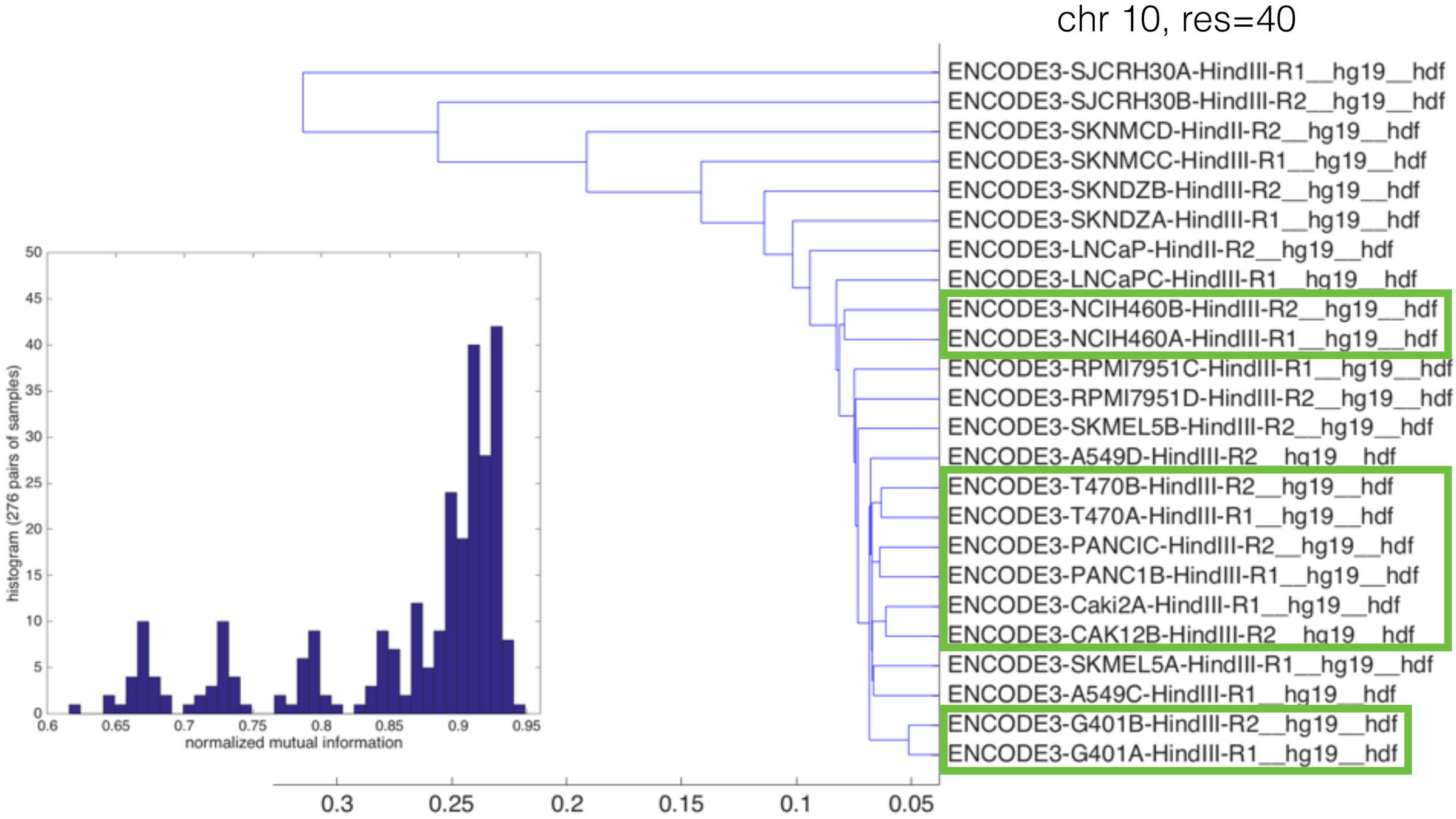
H3K27ac
H3K27me3
H3K36me3
H3K4me1
H3K4me3
H3K9me3



whole chromosome

chr. 10, bin size: 40kb

TADs across samples



Network provides a system-wide perspective to Hi-C data

- Identifying multi-scale topological domains based on network modularity detection

- A network framework to examine how spatial organization of genes shape their expression pattern

A mapping between 2 spaces

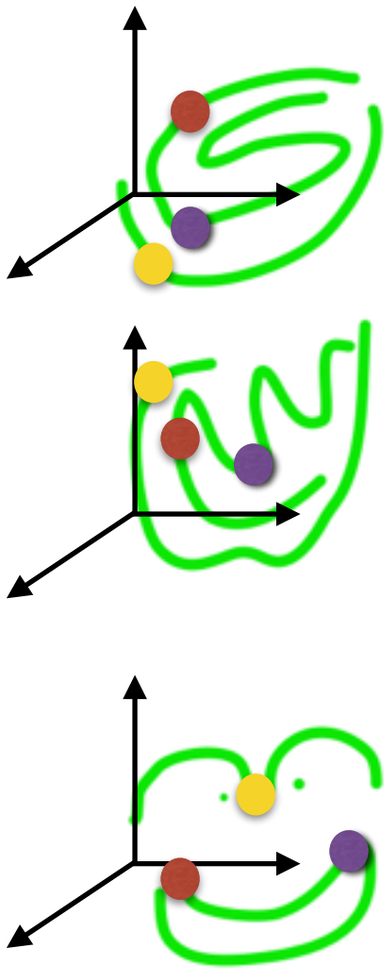
real physical space

abstract expression space

cell type 1

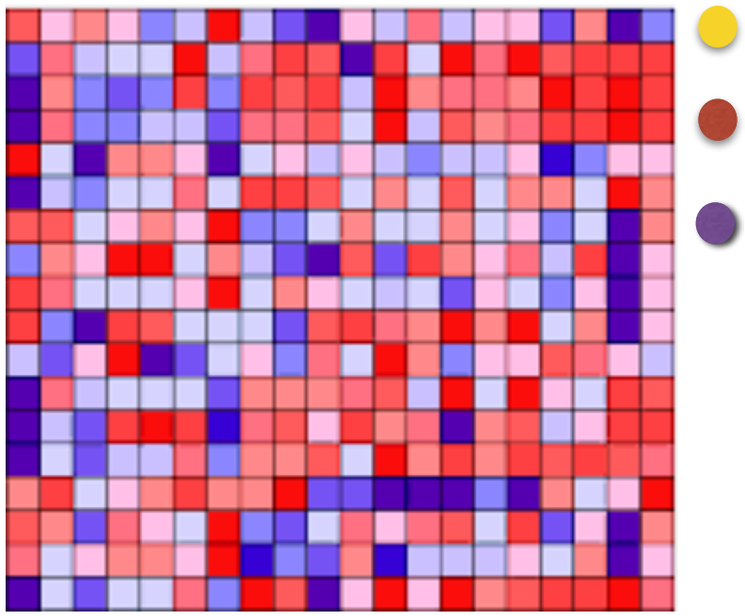
cell type 2

cell type 3



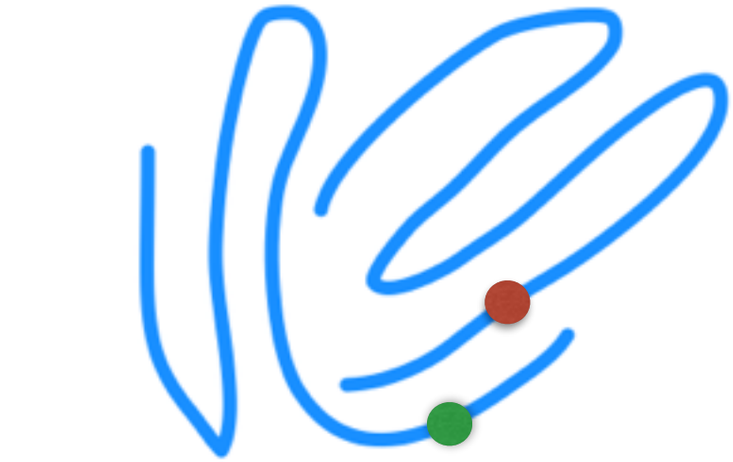
?

cell types: 1, 2, 3 …



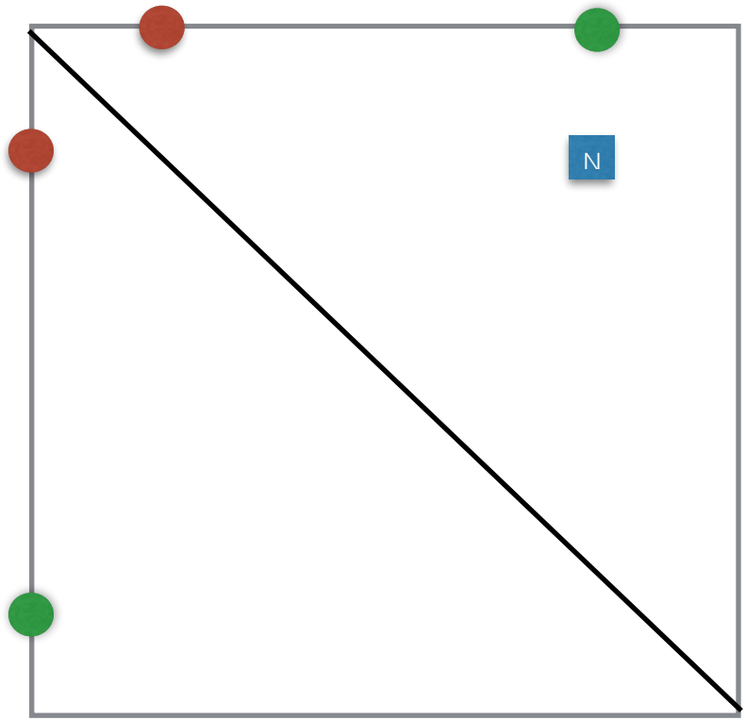
Gene

A simple construction: Gene-Gene Proximity Network



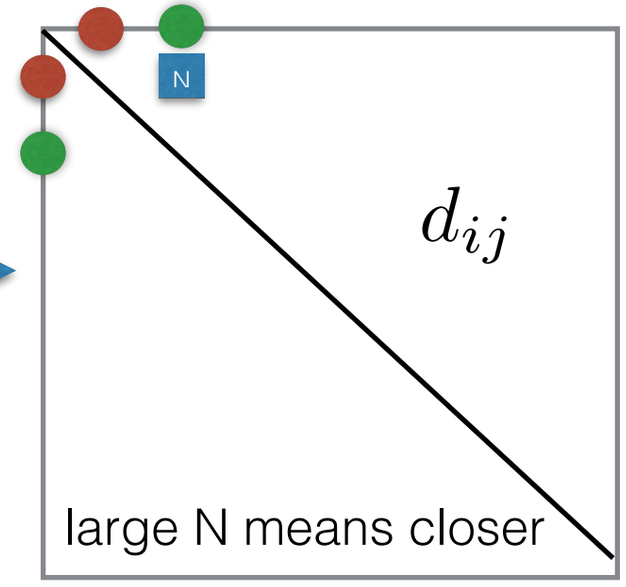
2 genes
(interchromosomal/intrachromosomal)

Hi-C contact matrix



Genomics coordinates
100kb resolution, ICED

Gene-Gene proximity matrix/network

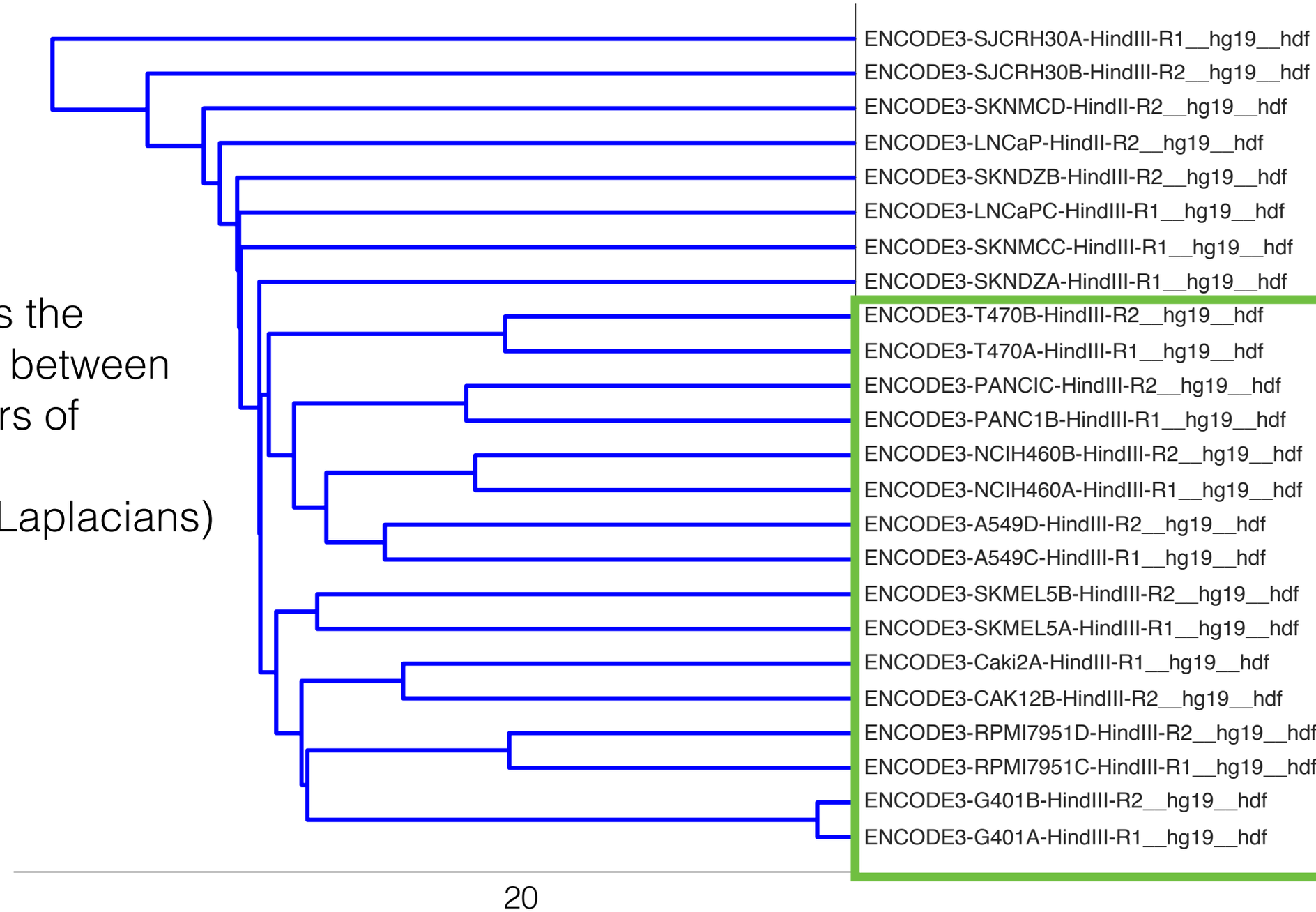


all genes

Example: A549
19100 genes
14% of gene pairs
are connected

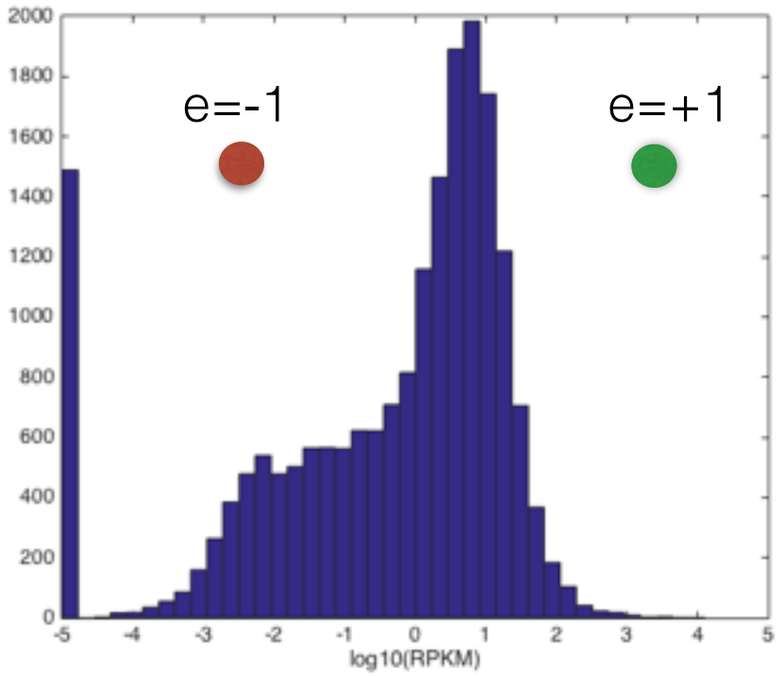
Gene-Gene Proximity Network across samples

Distance defined as the Euclidean distance between leading eigenvectors of corresponding diffusion matrices (Laplacians)

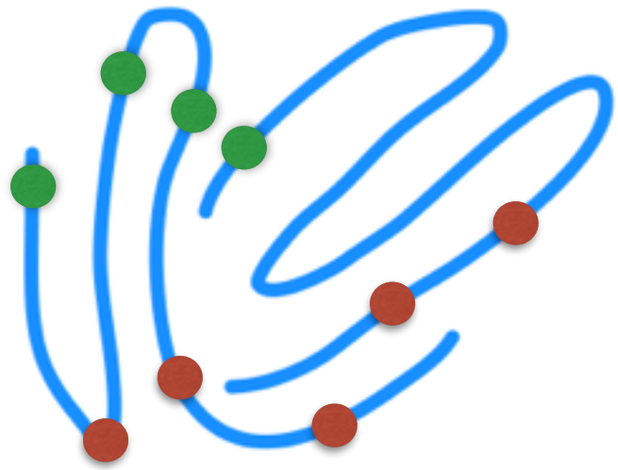


Gene-Gene proximity versus Gene-Gene expression

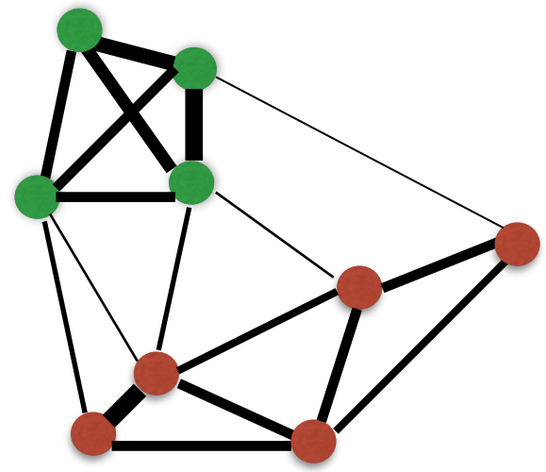
expression pattern of A549



spatial structure of A549



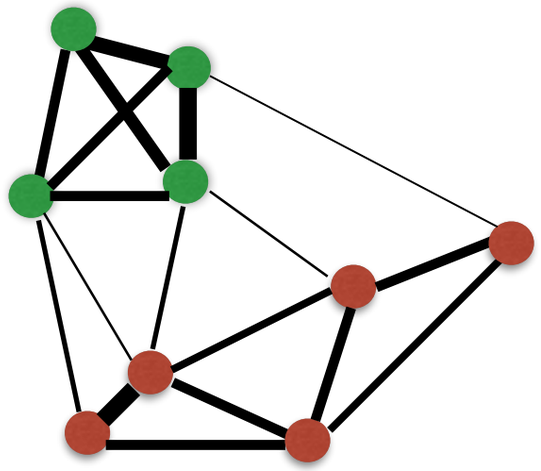
proximity network of A549



Graph partition (bisection) problem

Consider a graph G = (V, E), where V denotes the set of n vertices and E the set of edges. The objective is to partition G into k (k=2) components while minimizing the weights of the edges between separate components.

proximity network of A549

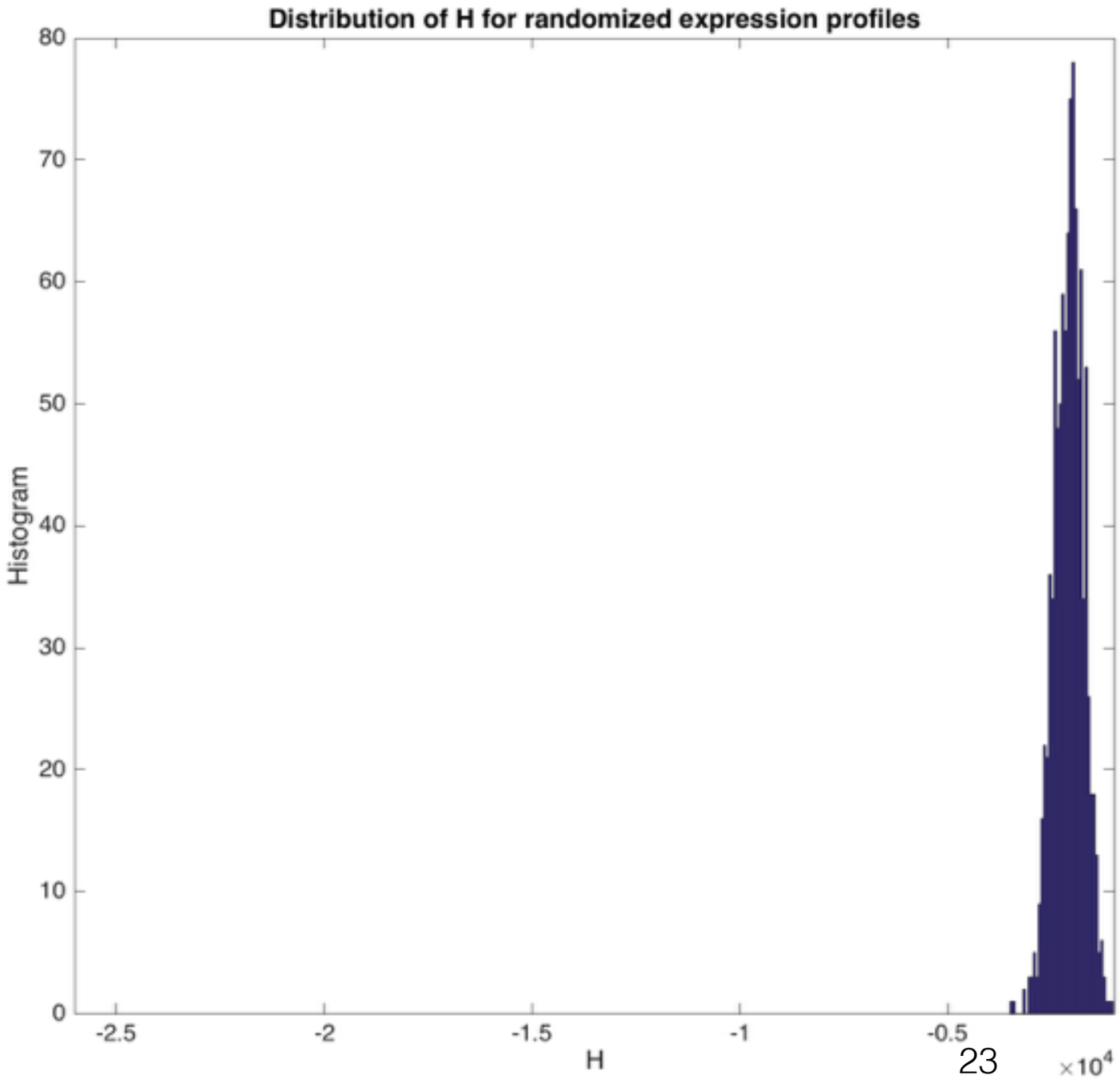


$$H = -\sum_{ij} d_{ij} e_i e_j$$

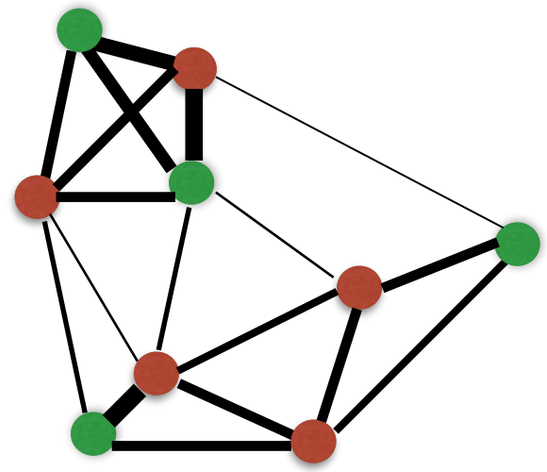
d is the weighted adjacency matrix and e=+1 or -1

a low energy state means co-expressed genes are co localized

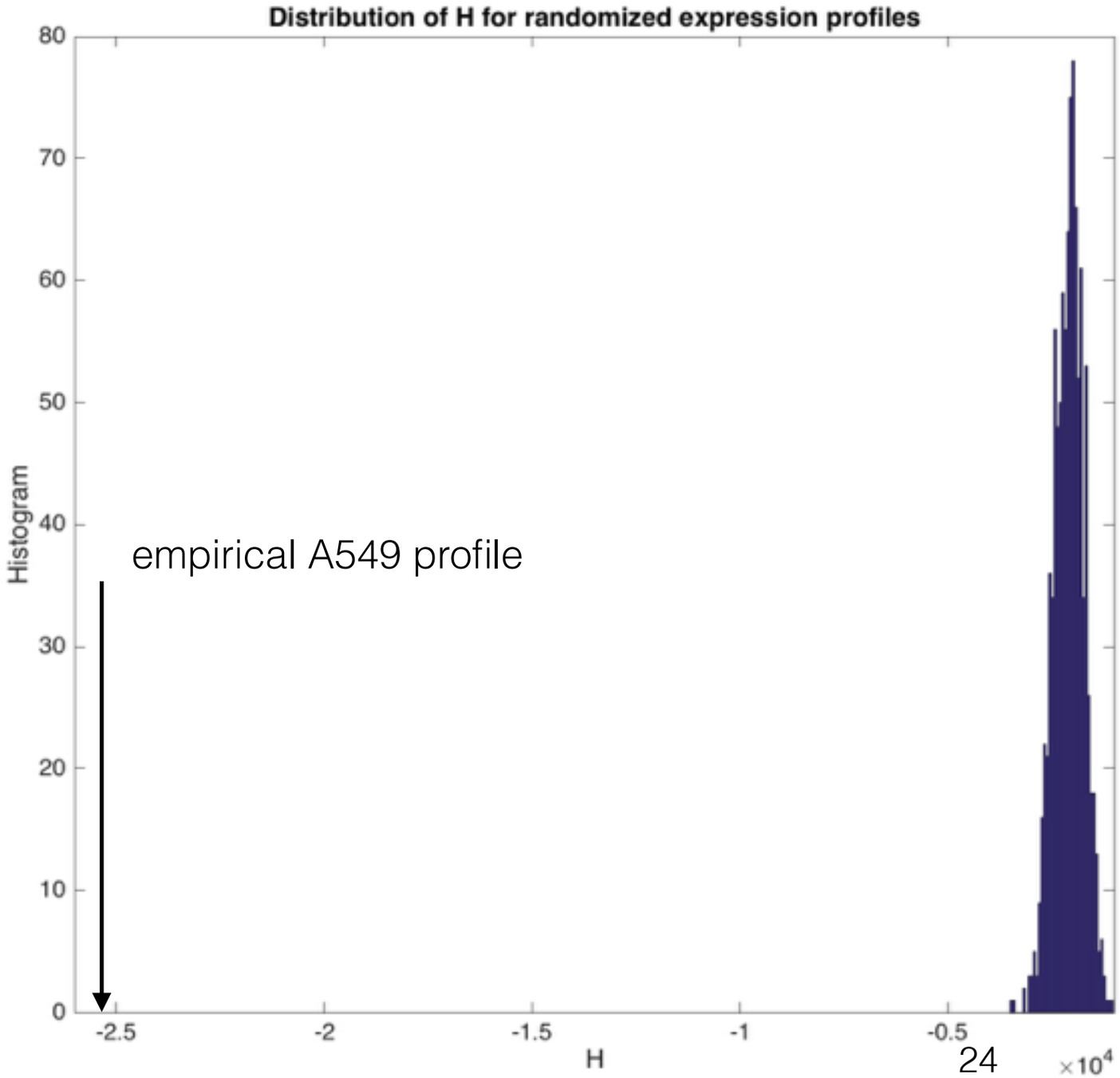
Gene-Gene proximity versus Gene-Gene expression



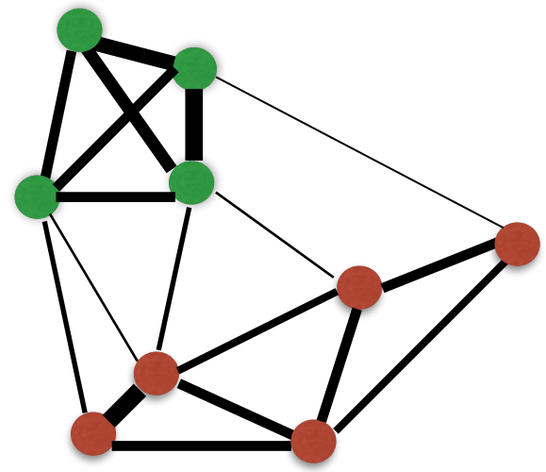
N nodes:
m is expressed, n is not



Gene-Gene proximity versus Gene-Gene expression

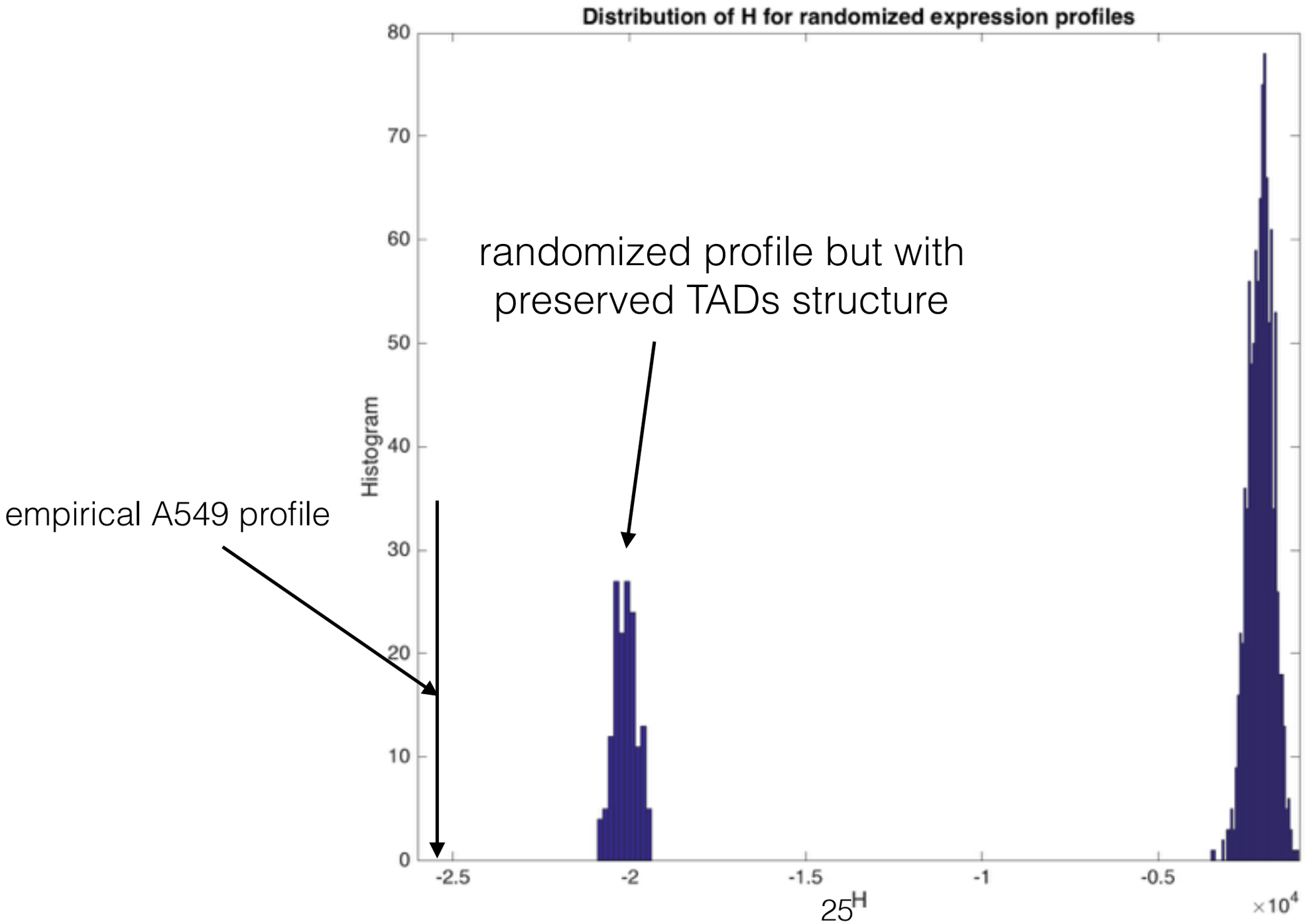


N nodes:
m is expressed, n is not



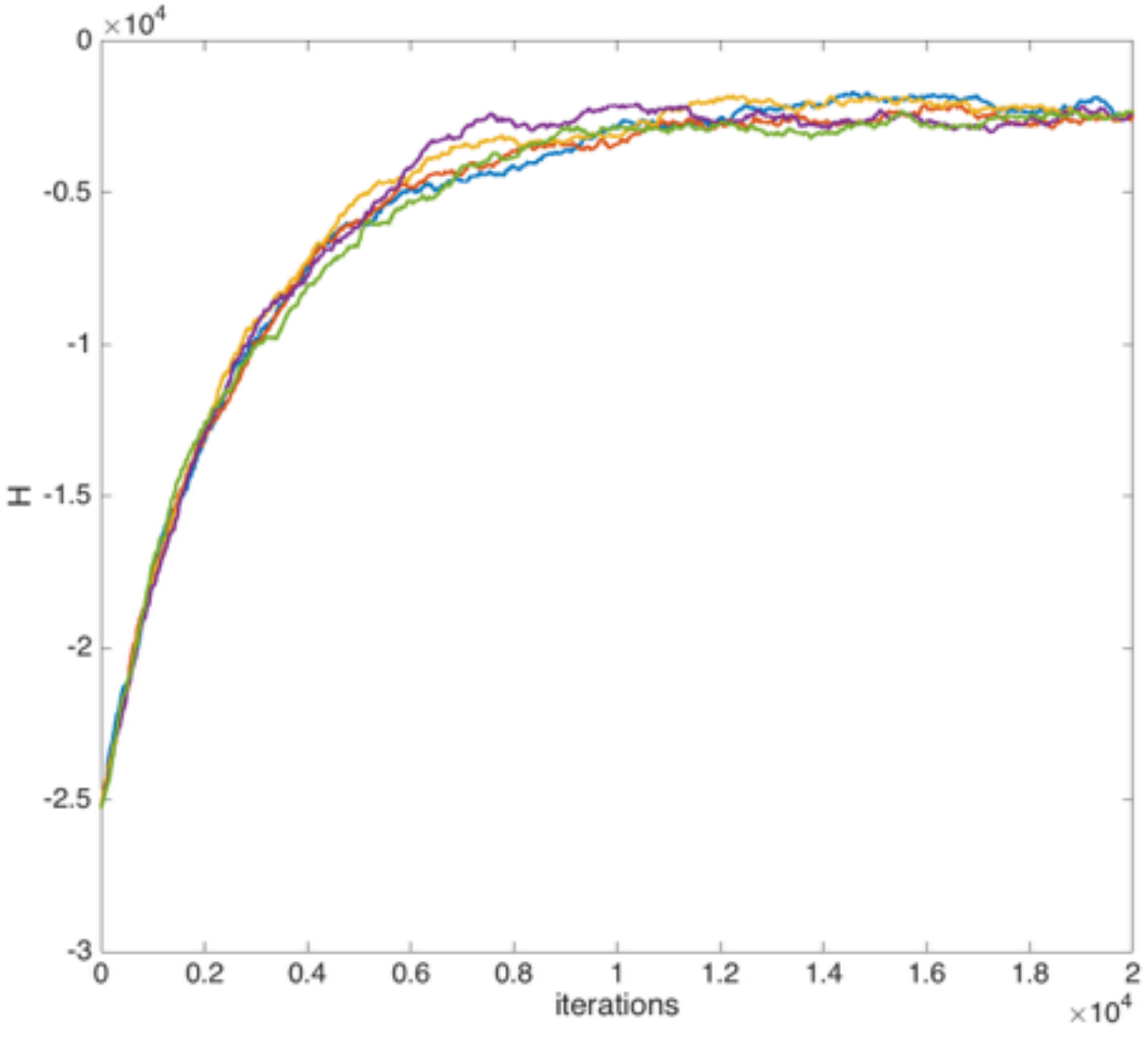
The spatial location of expressed genes are highly non-random.
May be it's too naive to compare with random - perform shuffling while preserving other genomics features

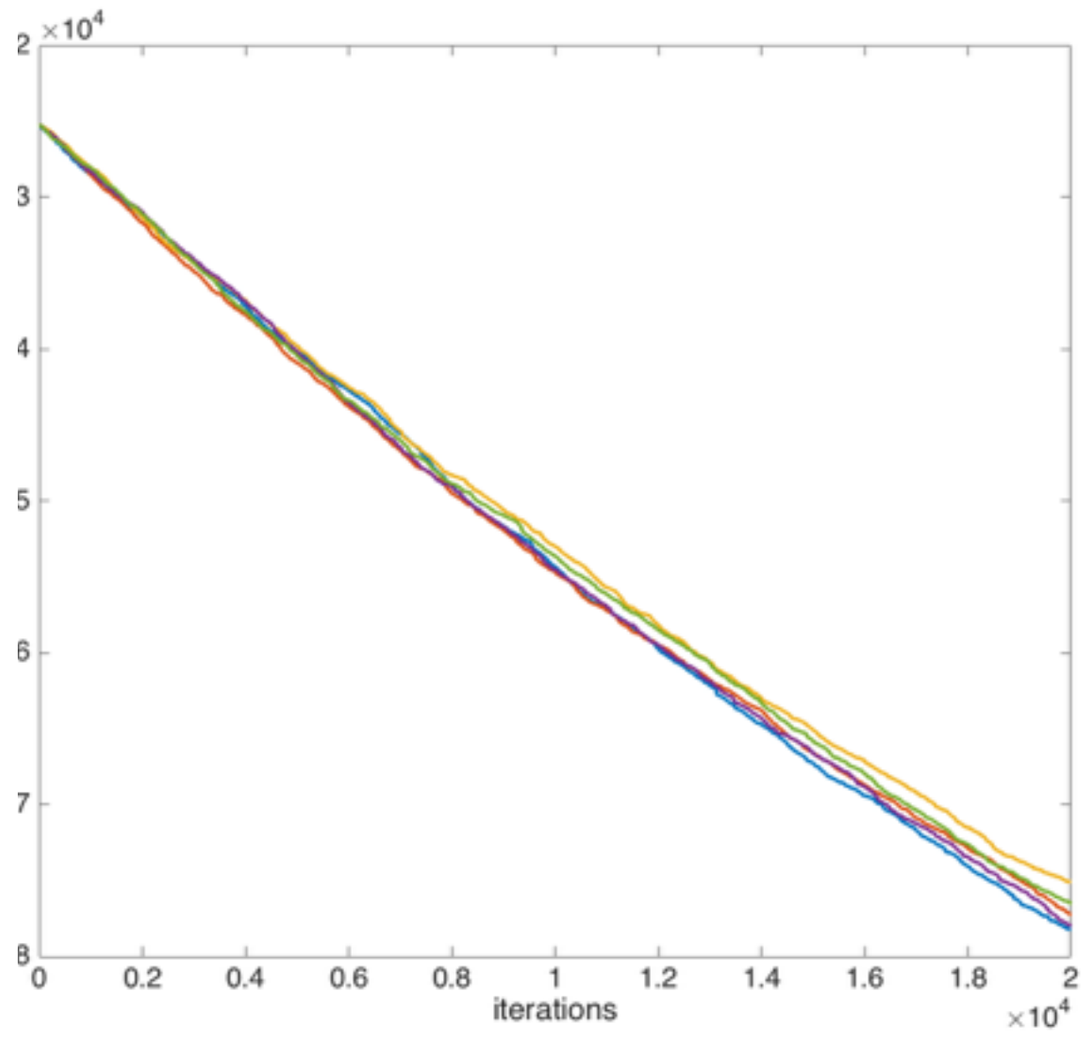
Effects of TADs



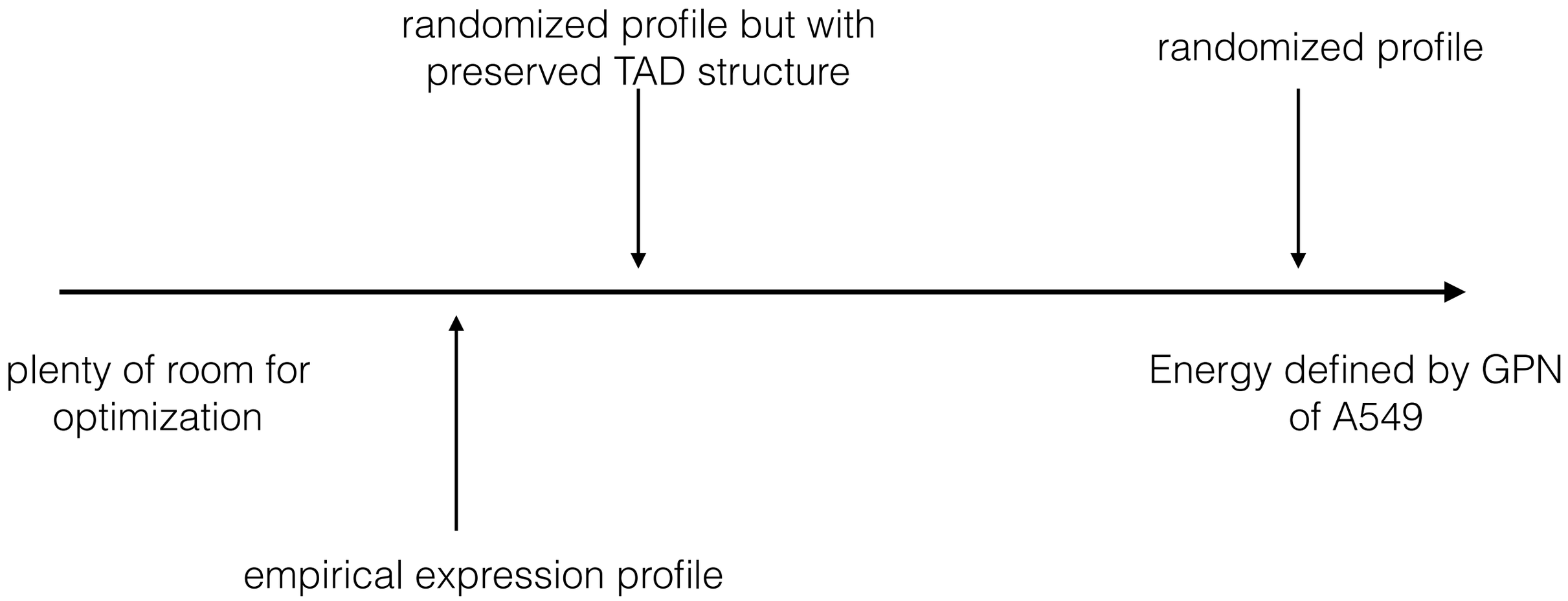
Is the expression profile optimal?

Given a spatial configuration, the observed expression profile has a much lower energy than random, but is it optimal?

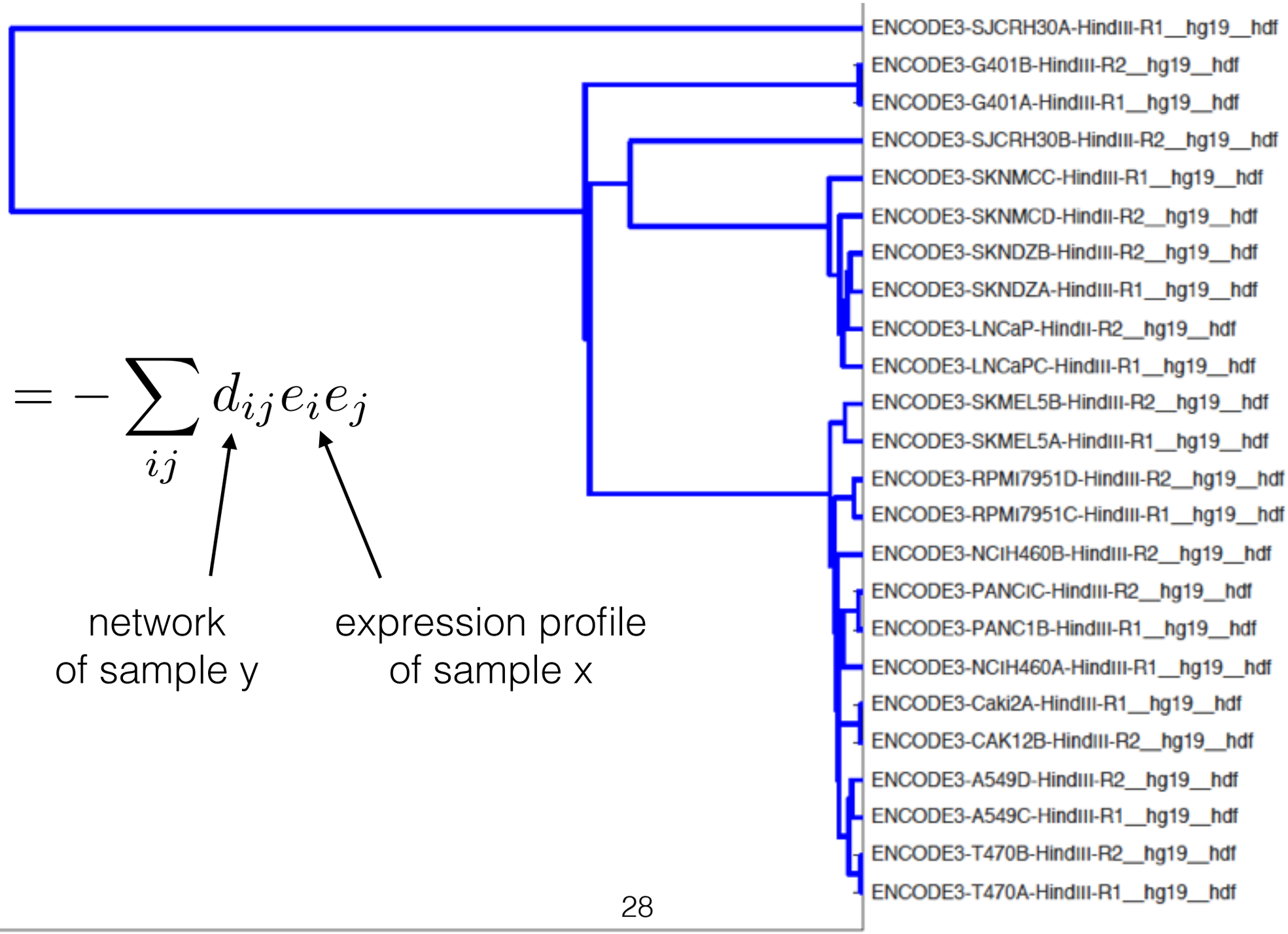




Is the expression profile optimal?



Matching expression patterns with Gene-Gene proximity in different samples



$$H = -\sum_{ij} d_{ij} e_i e_j$$

network of sample y

expression profile of sample x

Summary and In Progress

- Multi-scale TADs
    - developed an algorithm to detect TADs; TADs may exist in different length scales (hierarchical organization of genome: loop, sub-domains, TADs, compartments etc)
    - chromatin signatures of TADs in different resolutions
    - compare with existing algorithms
    - better null models, like a polymer model
- Gene-Gene proximity network
    - formulated the relationship between expression and spatial configuration as a graph partition problem
    - incorporate the targets of various transcription factors
    - more on comparison across cell lines, differential expression versus differential spatial configuration

Acknowledgement

- Gerstein lab
  - Anurag Sethi
  - Joel Rozowsky
  - Arif Harmanci
- Dekker lab for generating samples and pre-processing the data in 12 cell lines