Continue.
Hi-C related work

KKY
Oct, 2015

Recap

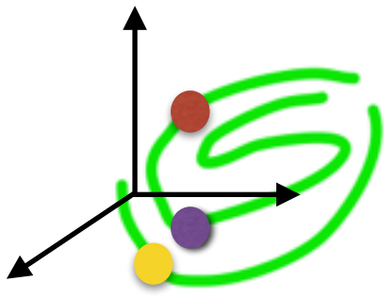
- How the spatial organization of genes shapes their expression patterns?

A mapping between 2 spaces

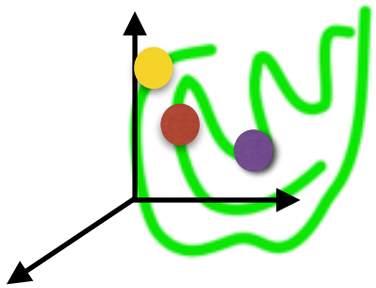
real physical space

abstract expression space

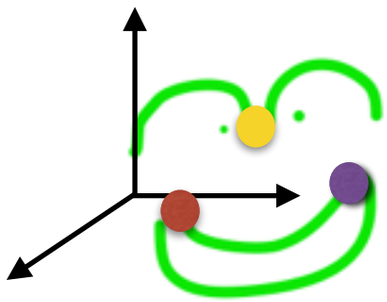
cell type 1



cell type 2

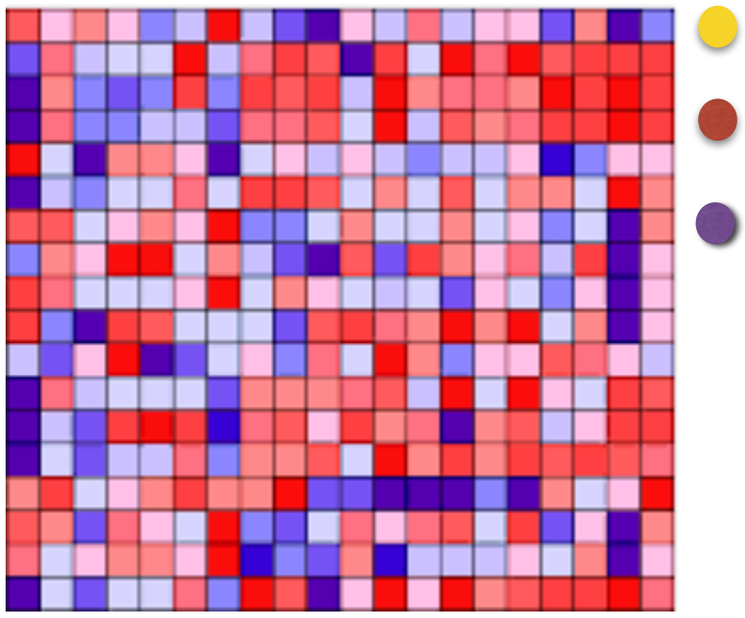


cell type 3



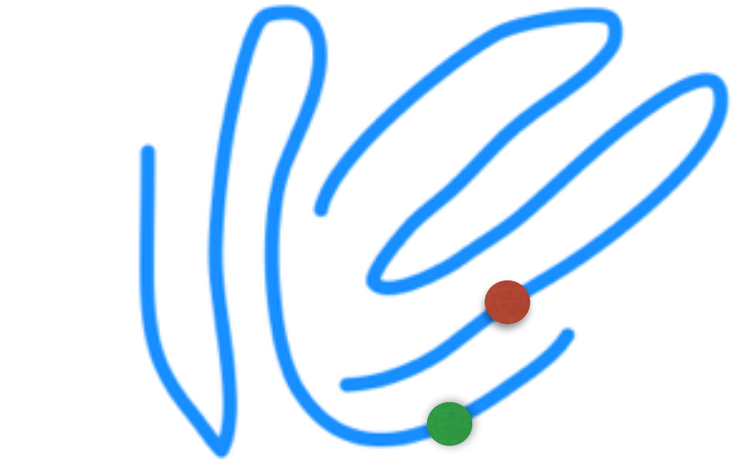
?

cell types: 1, 2, 3 …



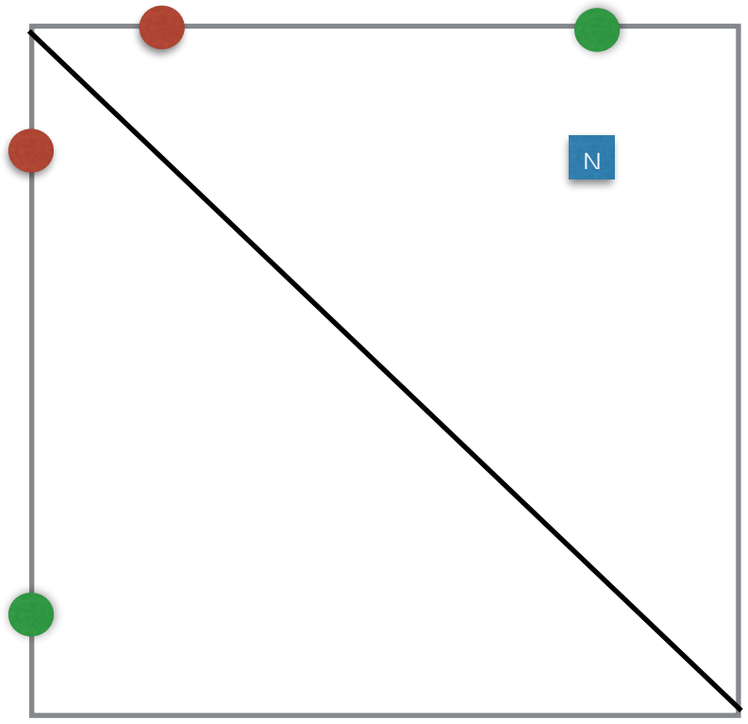
Gene

A simple construction: Gene-Gene Proximity Network



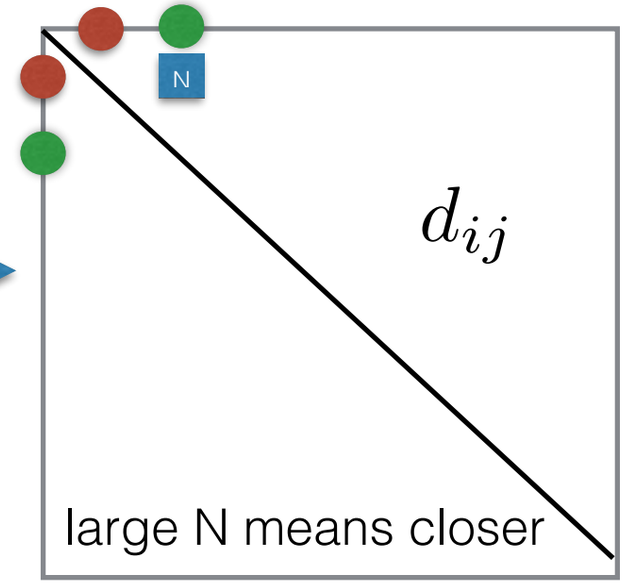
2 genes
(interchromosomal/intrachromosomal)

Hi-C contact matrix



Genomics coordinates
100kb resolution, ICED

Gene-Gene proximity matrix/network

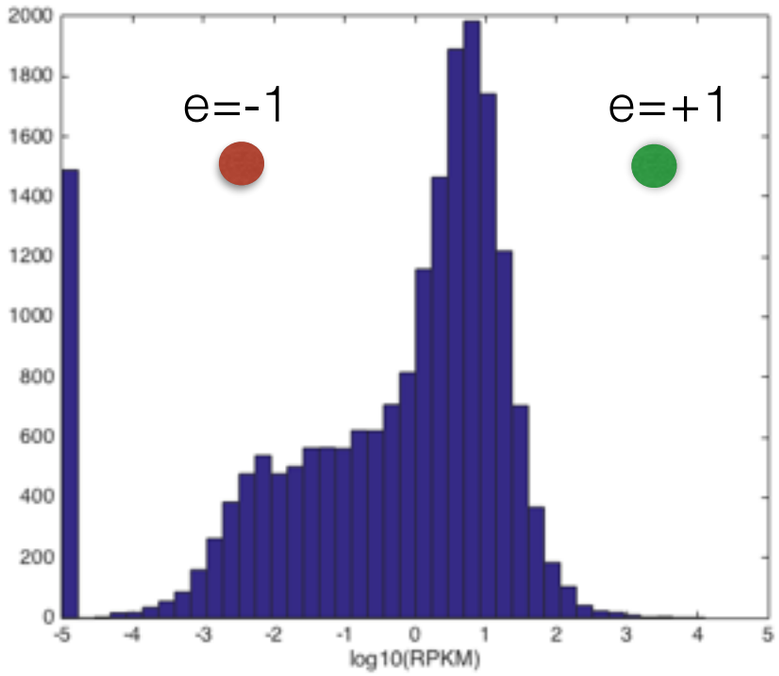


all genes

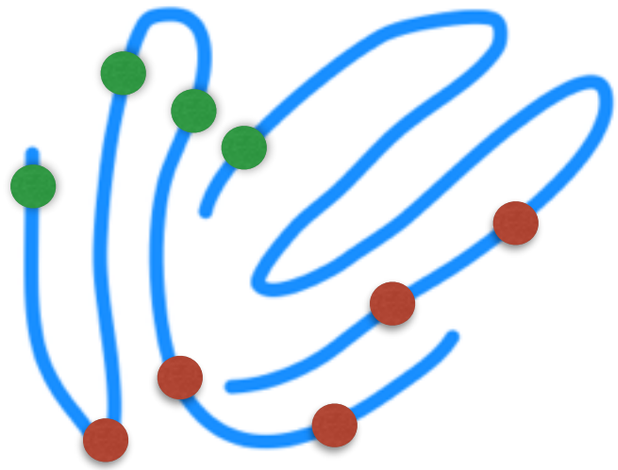
Example: A549
19100 genes
14% of gene pairs
are connected

Gene-Gene proximity versus Gene-Gene expression

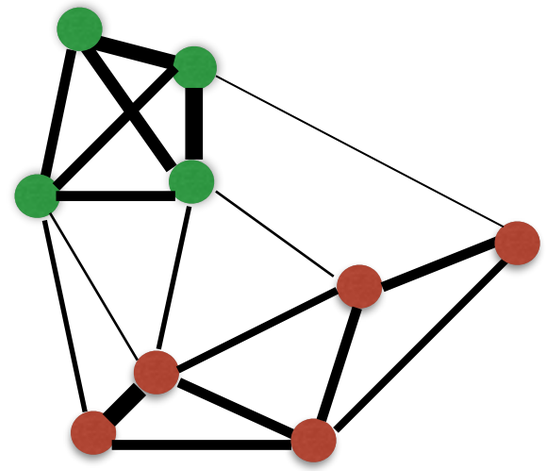
expression pattern of A549



spatial structure of A549



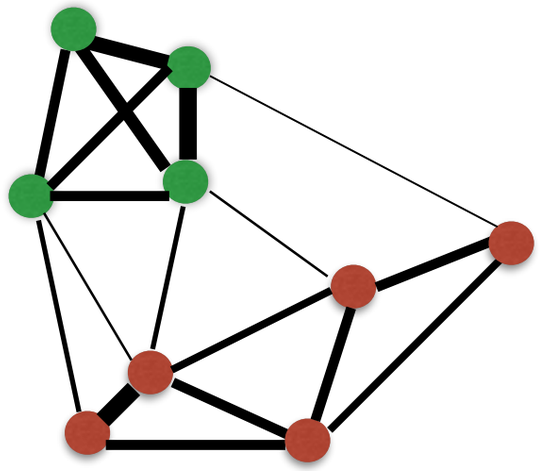
proximity network of A549



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



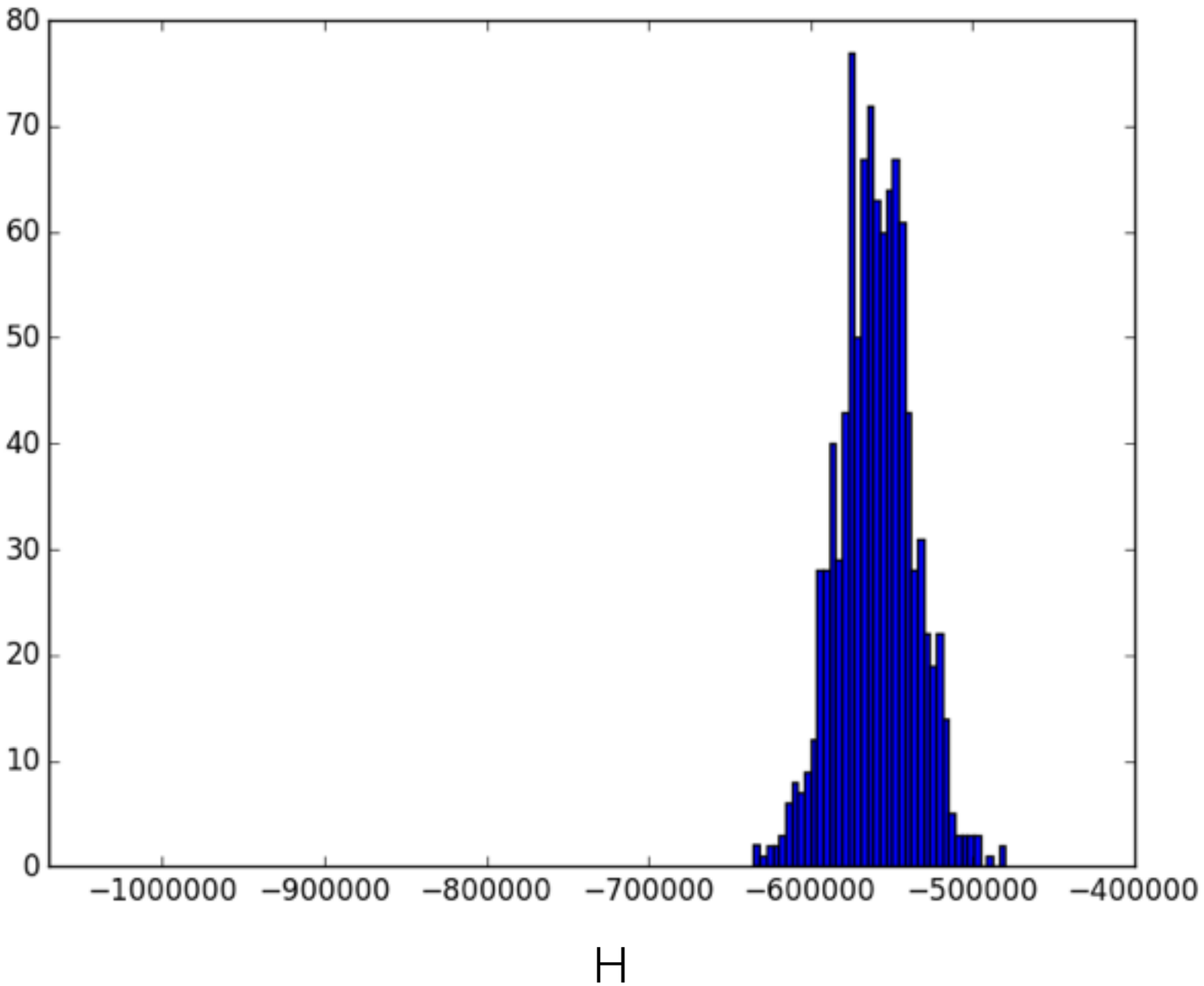
$$H = -\sum_{ij} d_{ij} e_i e_j$$

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

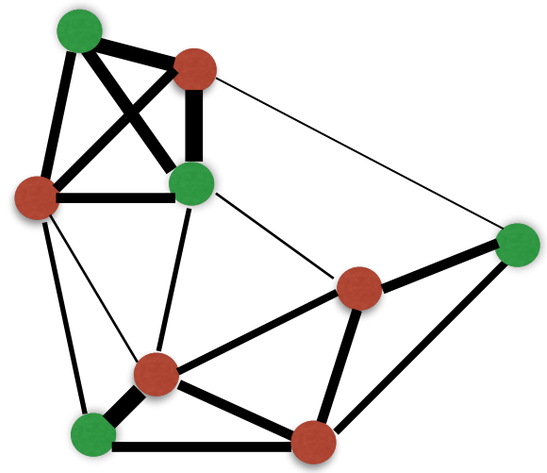
a low energy state means co-expressed genes are co localized

Gene-Gene proximity versus Gene-Gene expression

Distribution of H by shuffling the expression profile of A549

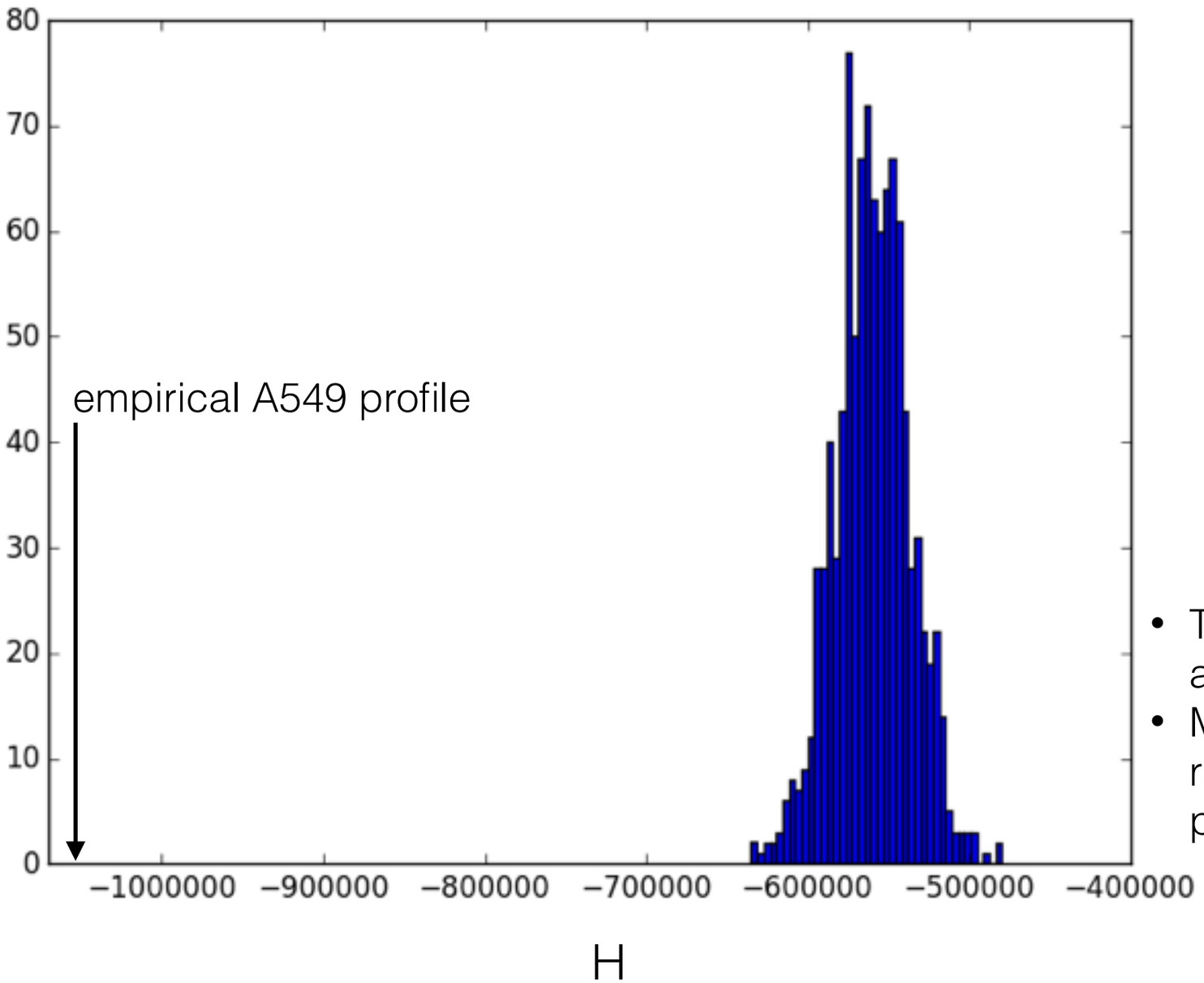


N nodes:
m is expressed, n is not

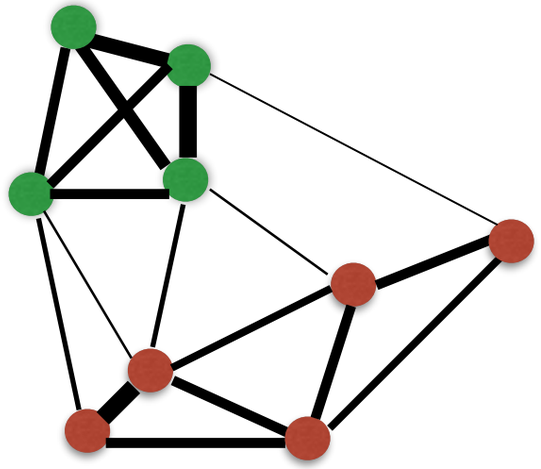


Gene-Gene proximity versus Gene-Gene expression

Distribution of H by shuffling the expression profile of A549



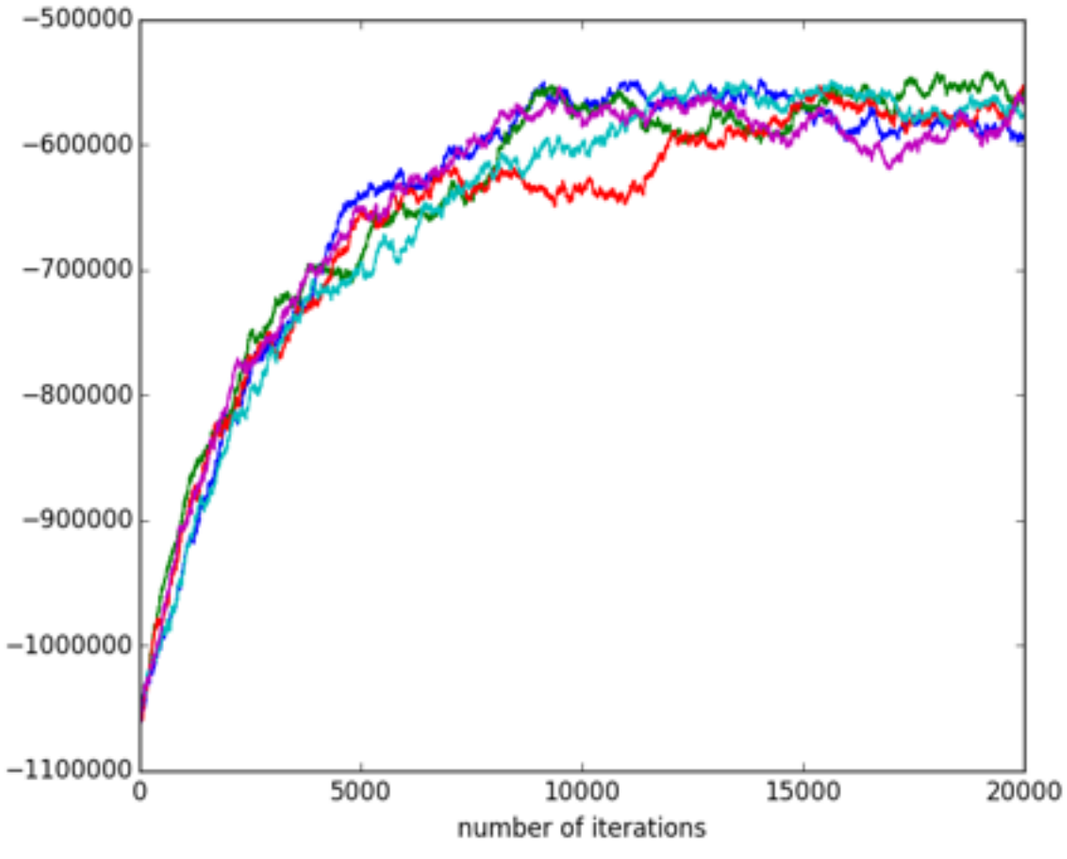
N nodes:
m is expressed, n is not



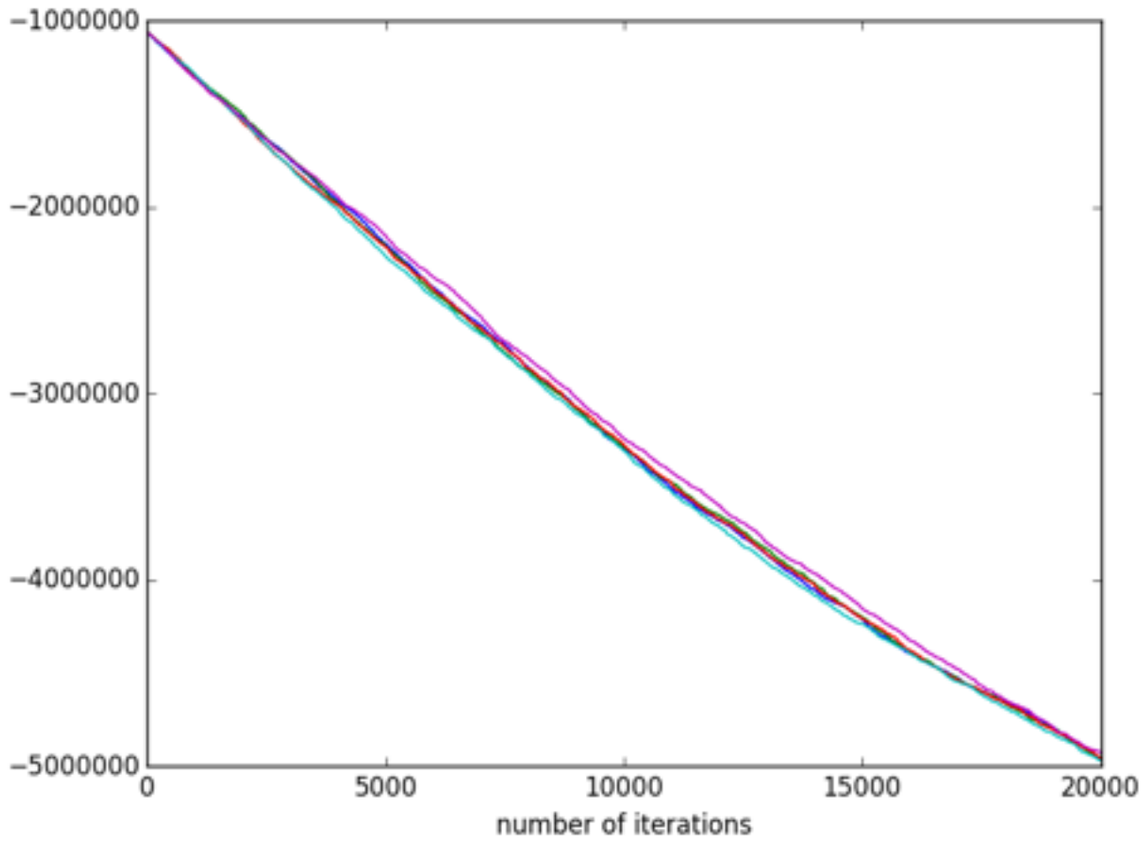
- The spatial location of expressed genes are highly non-random.
- May be it's too naive to compare with random - perform shuffling while preserving other genomics features

Is the expression profile optimal?

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



H



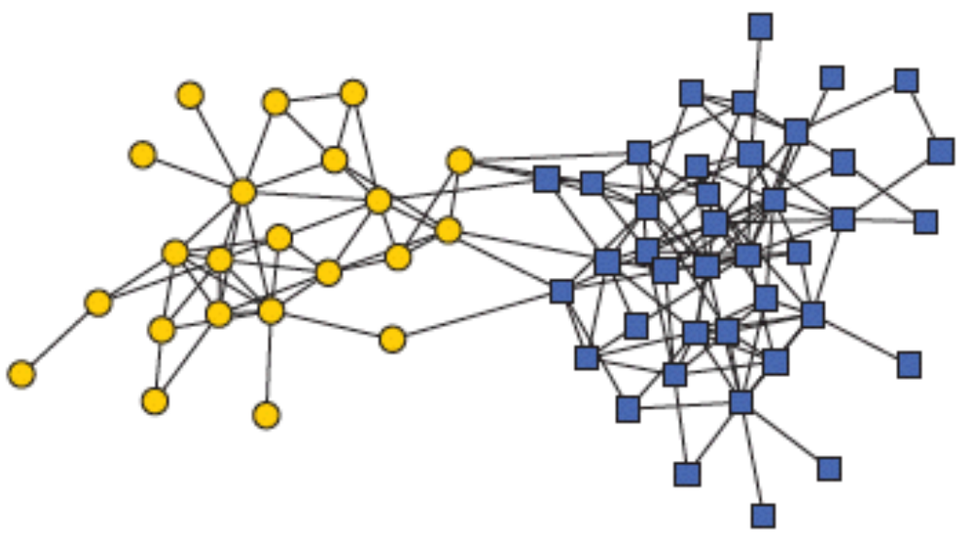
Updates

- Redid the analysis with different cell lines; redid the analysis with Hi-C data in the highest resolution (40kb); developed a better energy function

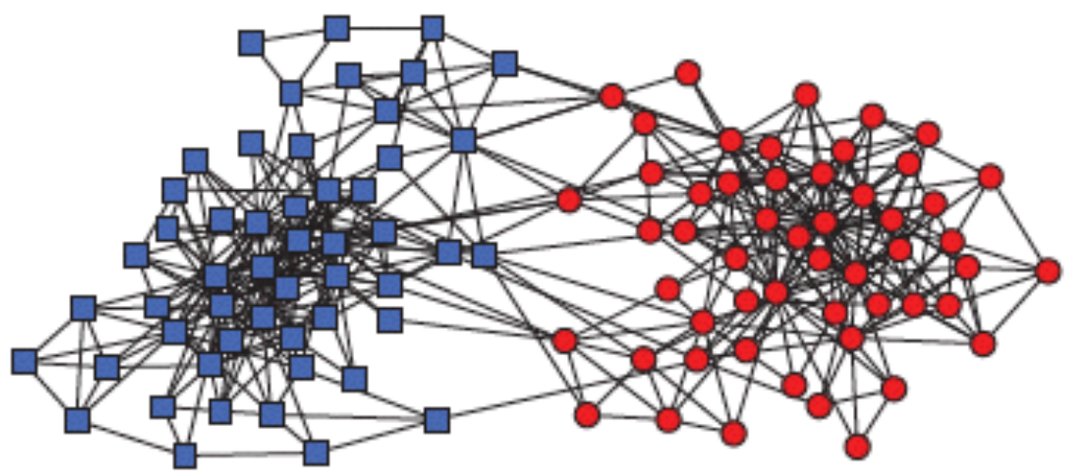
- bottleneck: incorporate the topological associated domains (TADs) into the expression analysis

- Identify TADs using network modularity

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

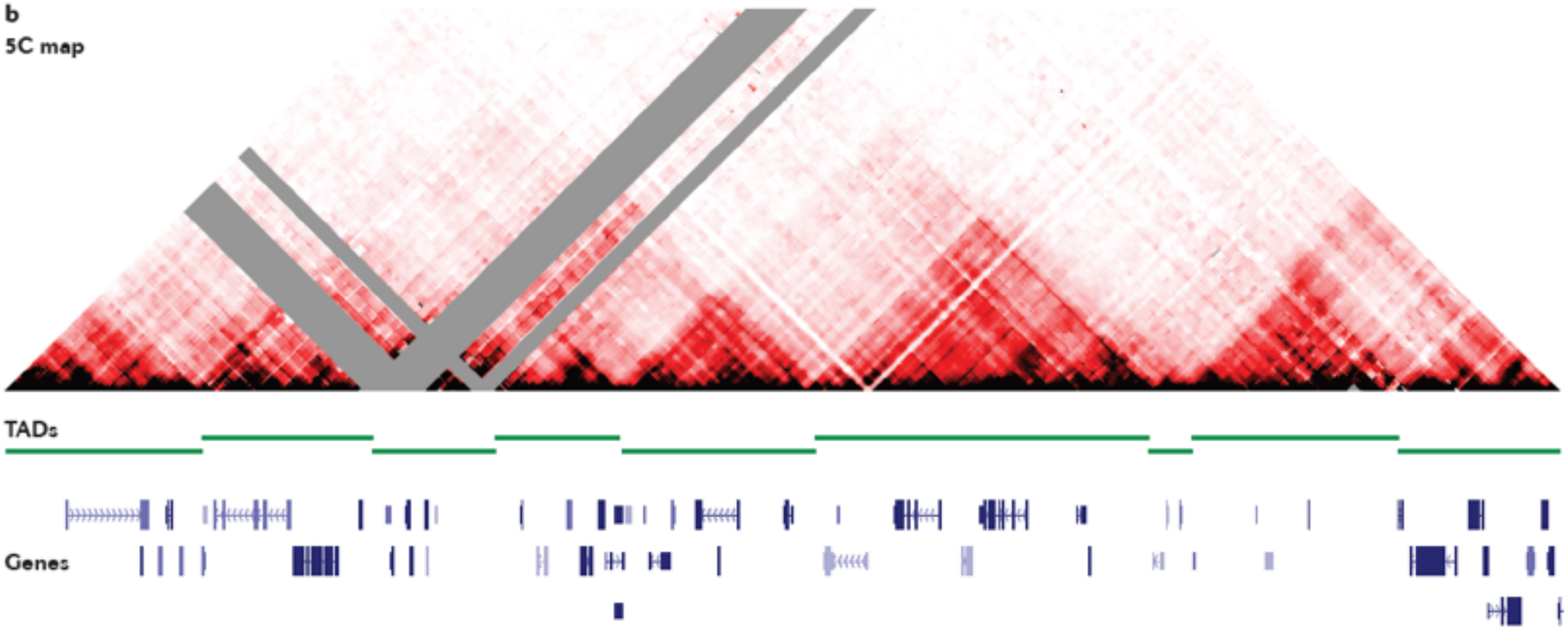
degree of i

whether or not i, j are in the same module

number of edges

expected number of edges between i and j

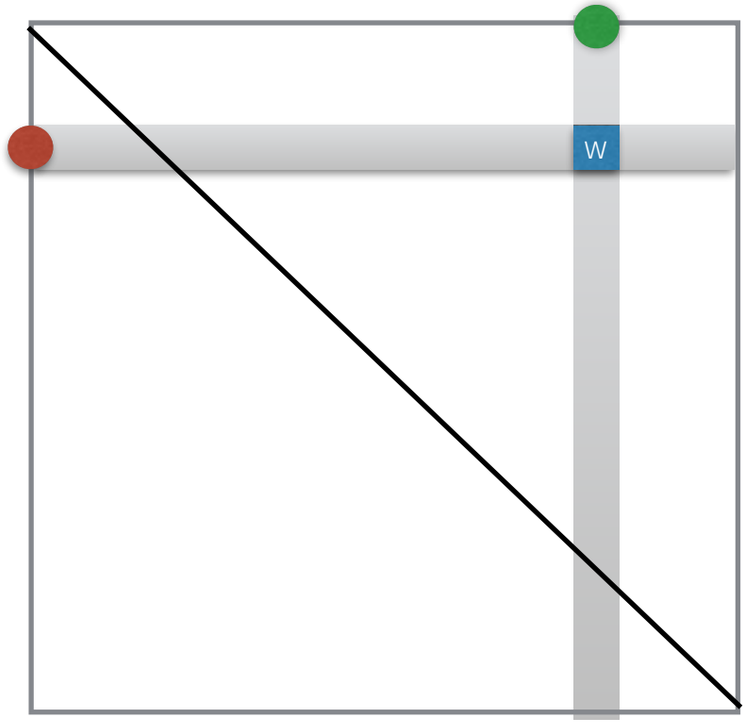
Topologically Associating Domains (TADs)



Dekker et al. Nat. Rev. Genetics 2013

Naive null model

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

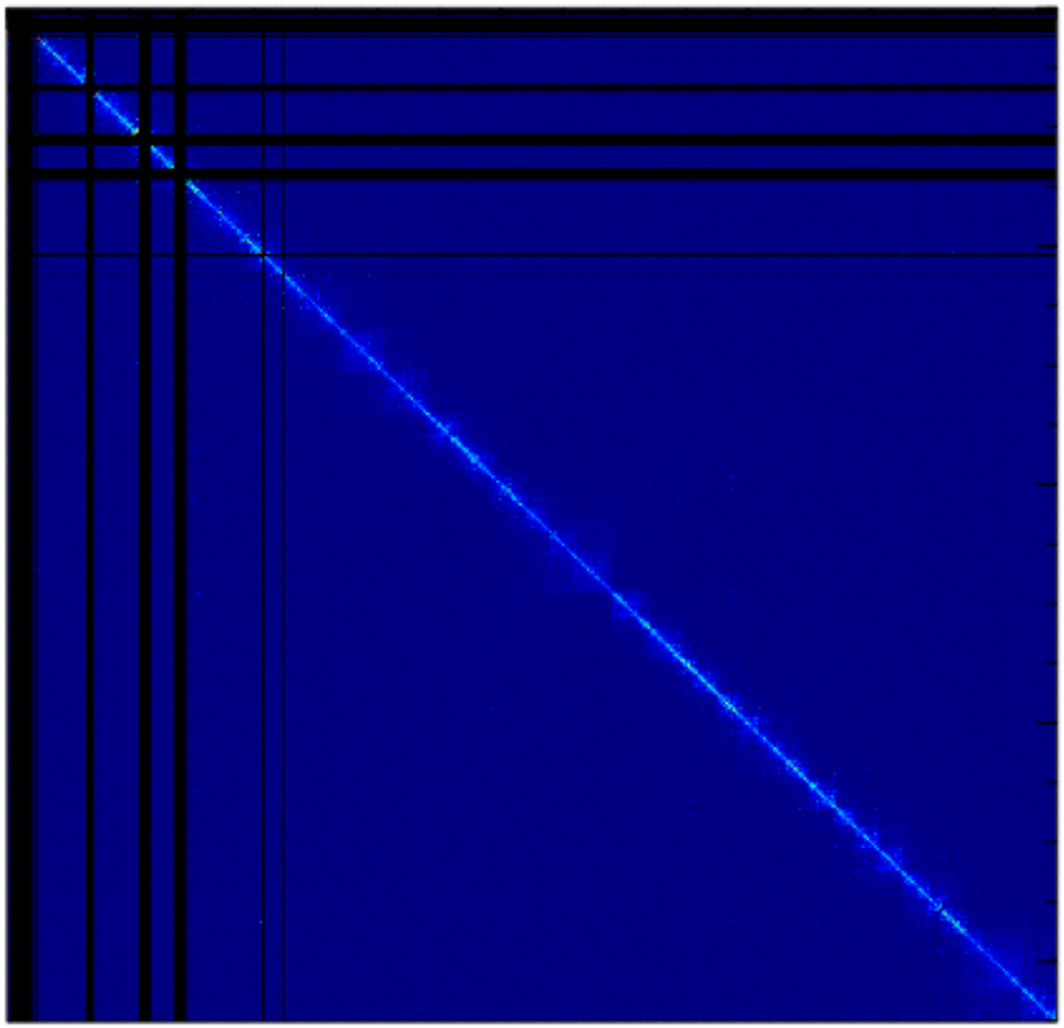
$$Q = \frac{1}{2N}\sum_{ij}(W_{ij} - \gamma\frac{aR_i aR_j}{2N})\delta_{\sigma_i\sigma_j}$$

resolution parameter

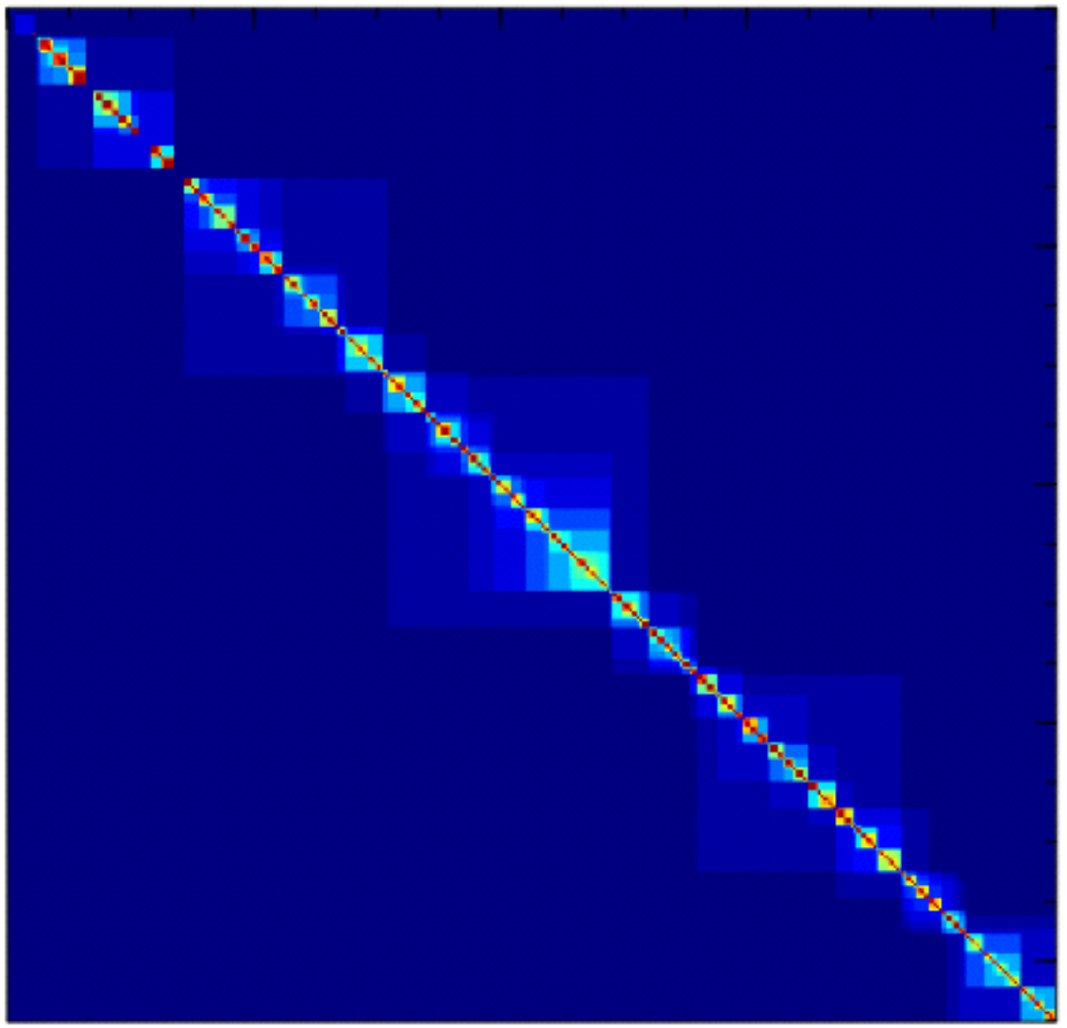
- An increase in gamma results in smaller modules
- An increase in gamma could be interpreted as focusing on the more statistically significant interactions (as compared to the null)
- Input: contact matrix (raw/iced) of the entire genome, or chromosome by chromosome (better in terms of TADs)

Examples

Hi-C contact (ICED)

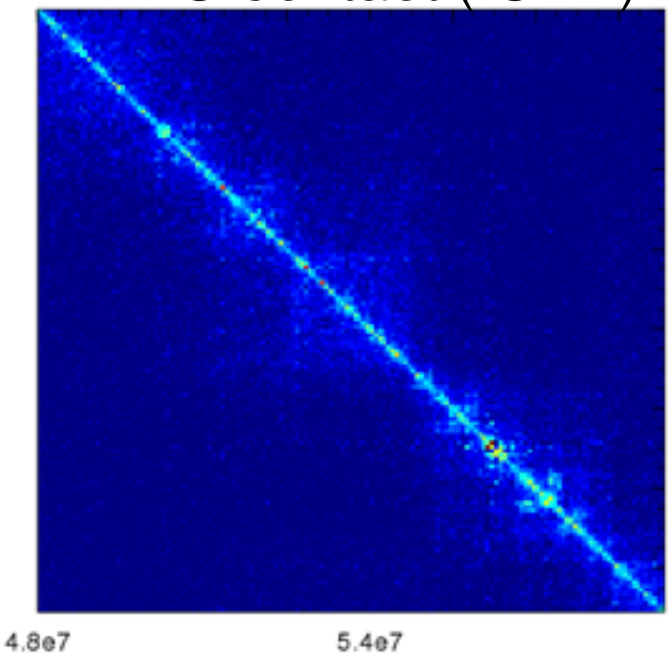


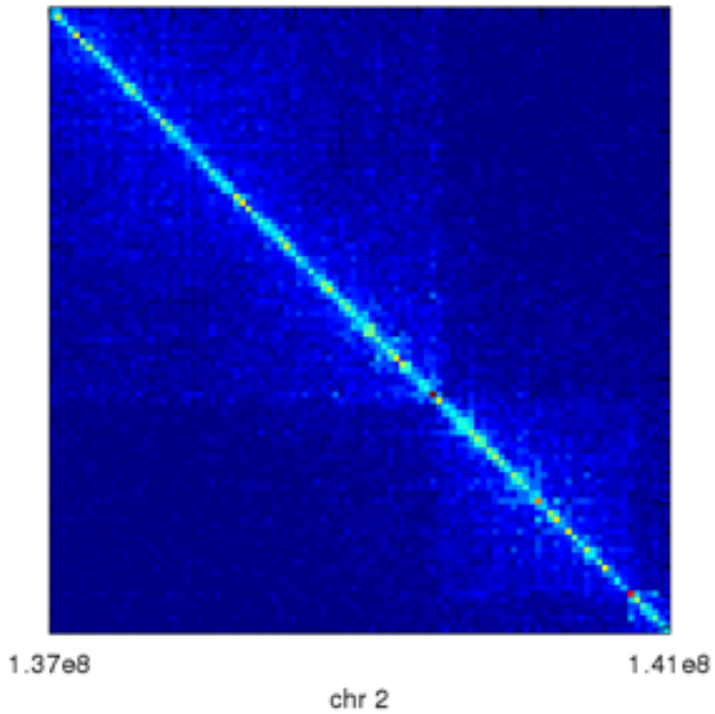
msTADs



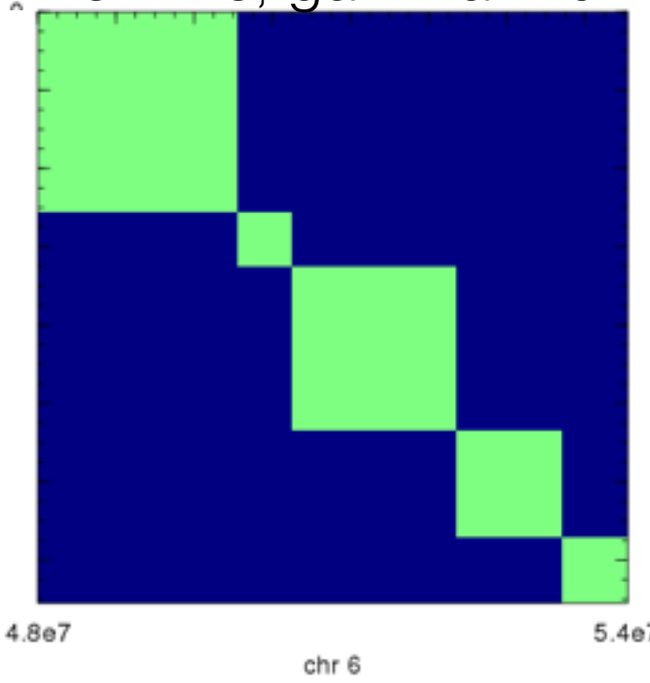
Examples (zoom in)

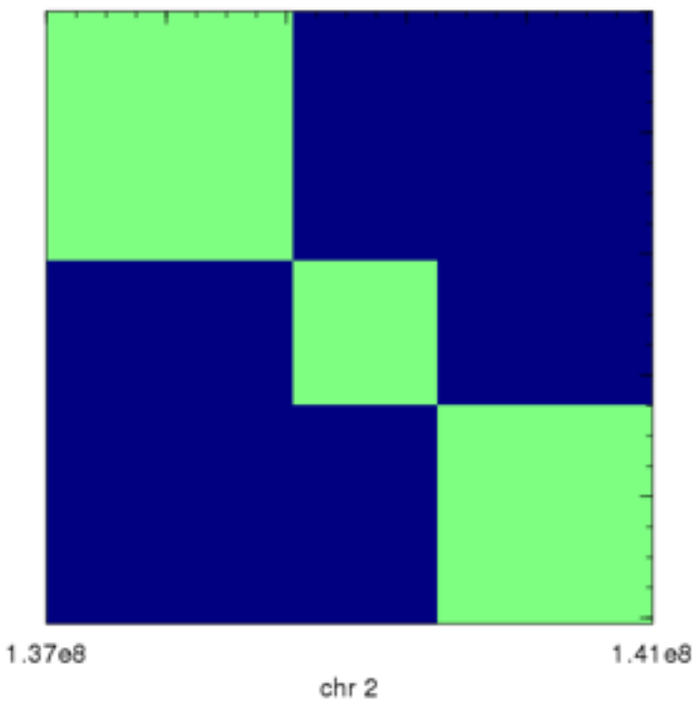
Hi-C contact (ICED)



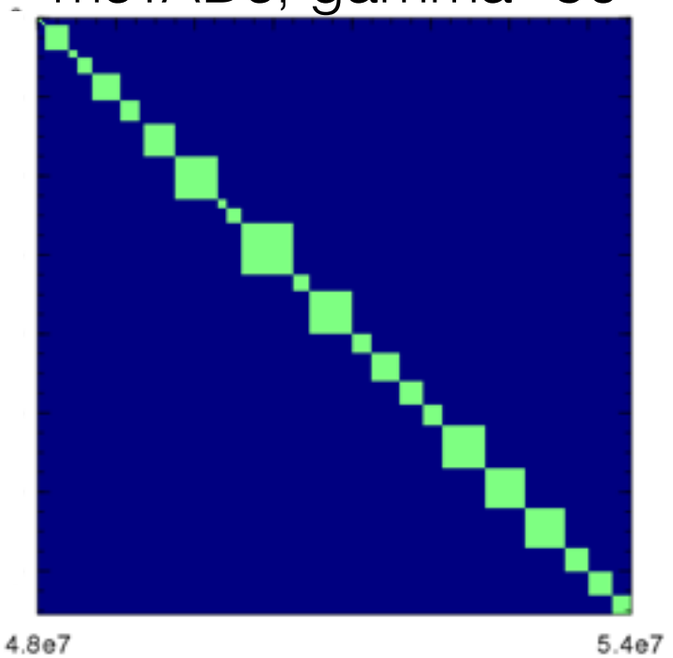


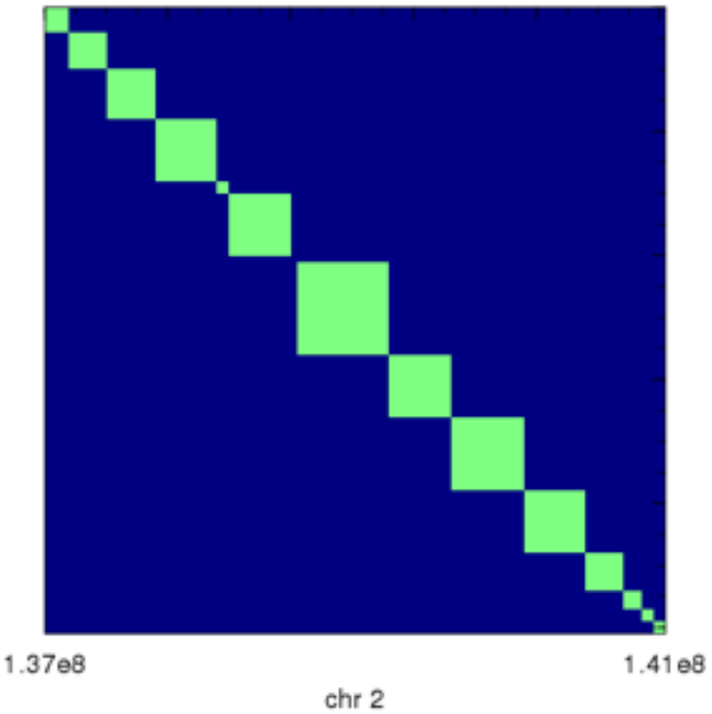
msTADs, gamma=10



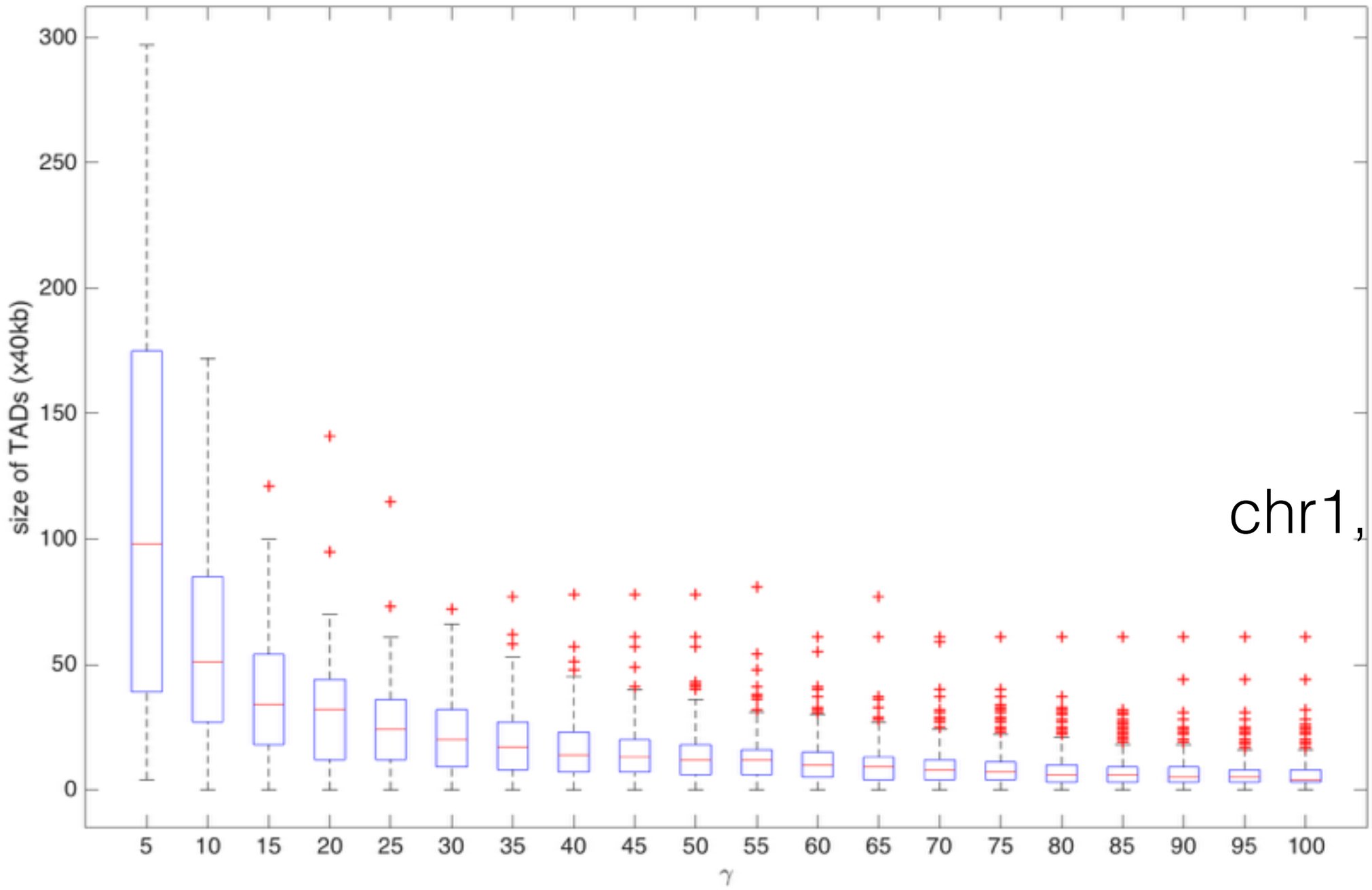


msTADs, gamma=50



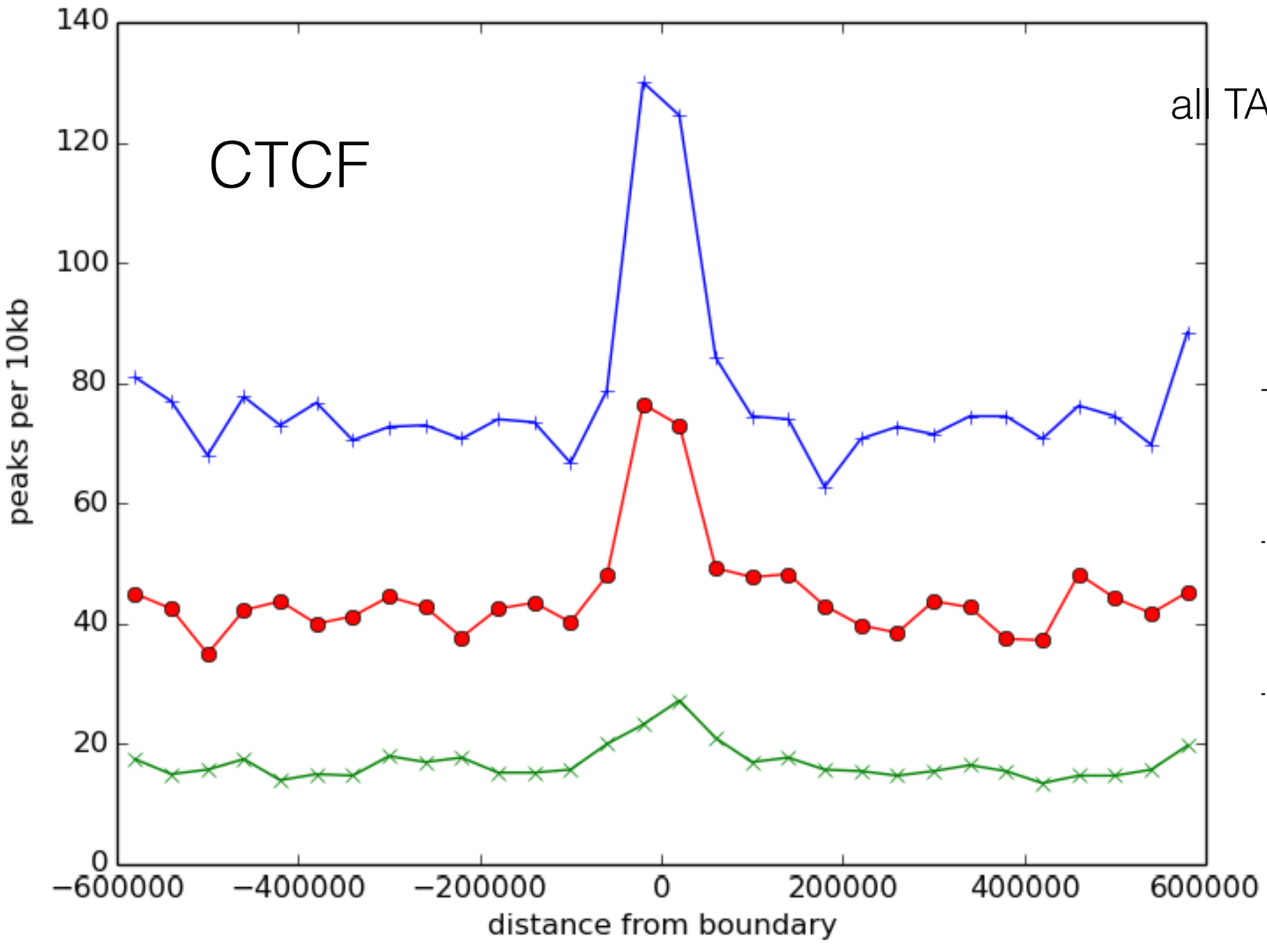


TADs size versus resolution



chr1, hESC

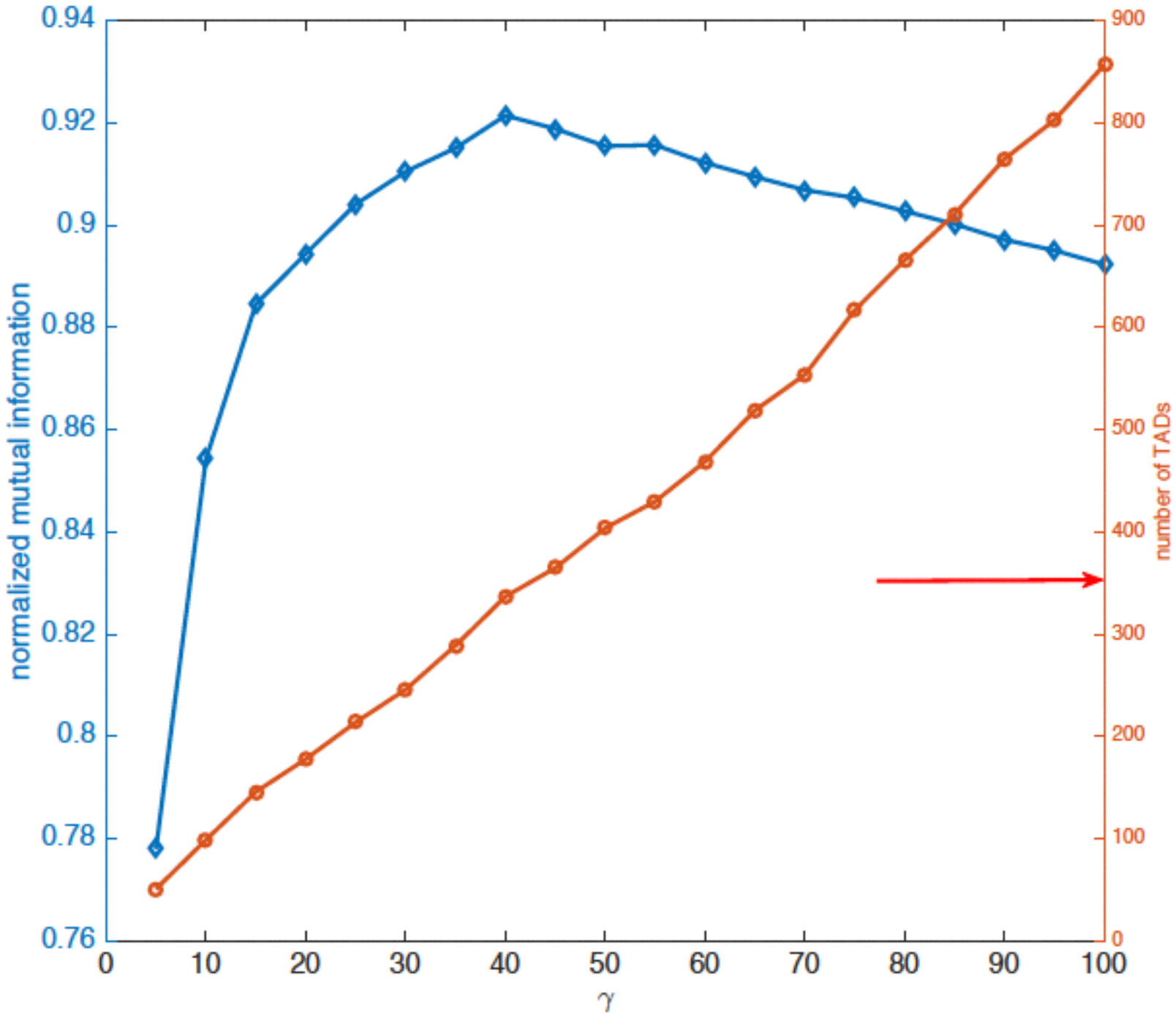
Boundaries between TADs



all TAD boundaries in chr1

gamma	num. of boundaries
10	102
30	301
50	578

Comparison with HMM method



based on
chr 1 of hESC

Discussion

- Developed an alternate approach to identify TADs from Hi-C data
- Results are comparable to conventional method (not sure if it is better, lack of gold standard)
- Novelty: multiple-resolution. How to make sense? multiple-scale chromatin states? MUSIC?
- across cell lines