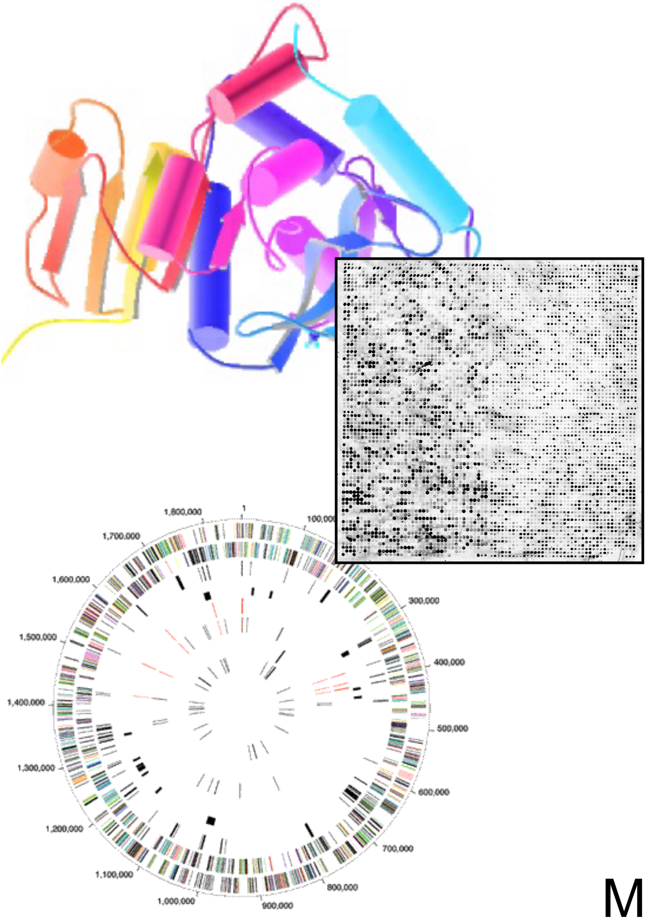
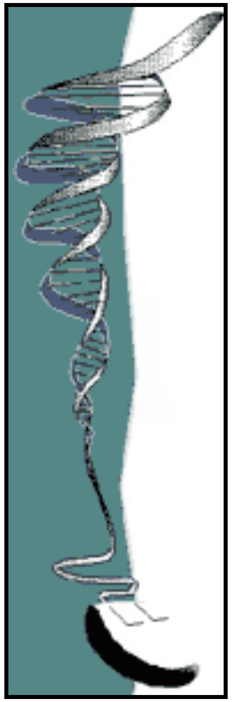
Biomedical Data Science
Hi-C Analysis





Mark Gerstein, Yale University
gersteinlab.org/courses/452
(last edit in Spring '19, pack #12)

Hi-C analysis illustrates much of the material in the class

- Provides an illustration of
  - How machine learning functions to make sense of large, complex datasets
  - Network topology
  - Aggregation plots
  - Spectral methods (SVD)

- Illustrates the evolution of the problem of annotating active & repressed regions in the genome
  - Original formulation in terms of "peak calling" on the linear genome
  - Revision of the original work, now at multi-scale
  - Recent radical change: now thinking of the genome as a 3D folded molecule

3D organization of genome

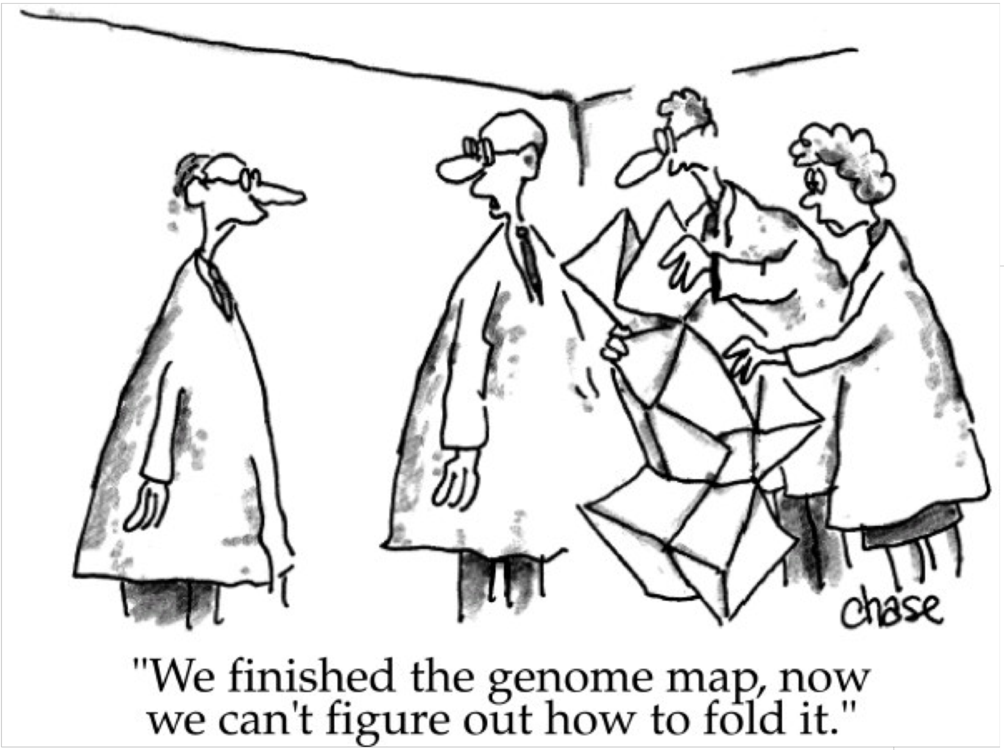

chase
"We finished the genome map, now we can't figure out how to fold it."

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

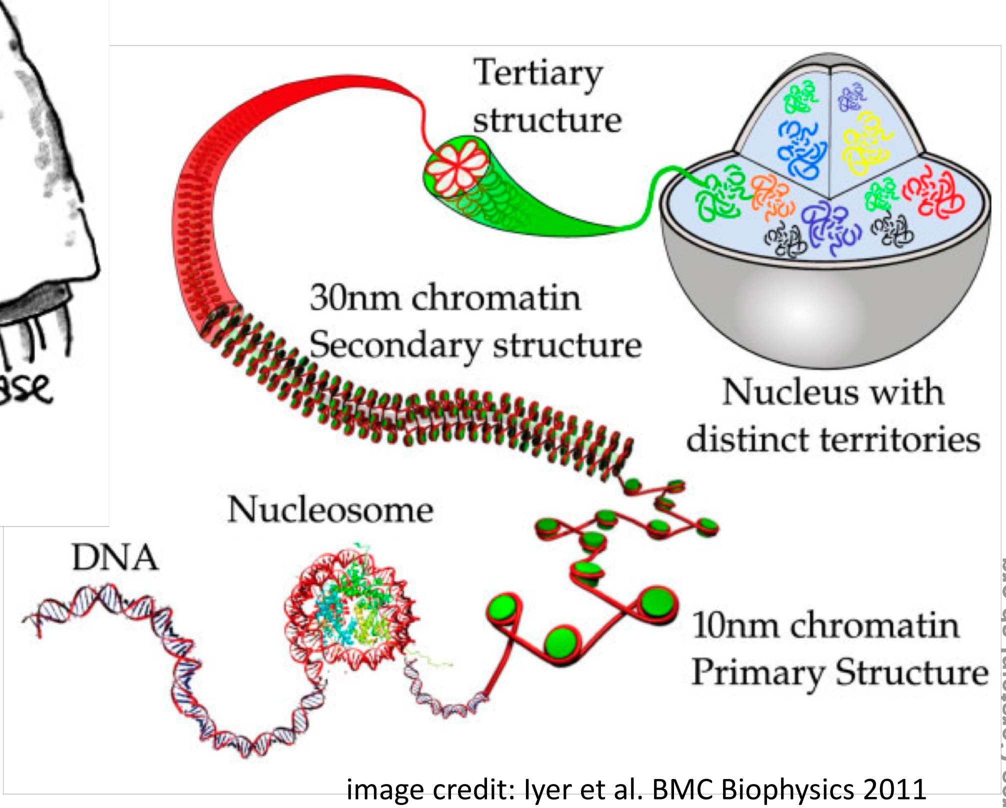
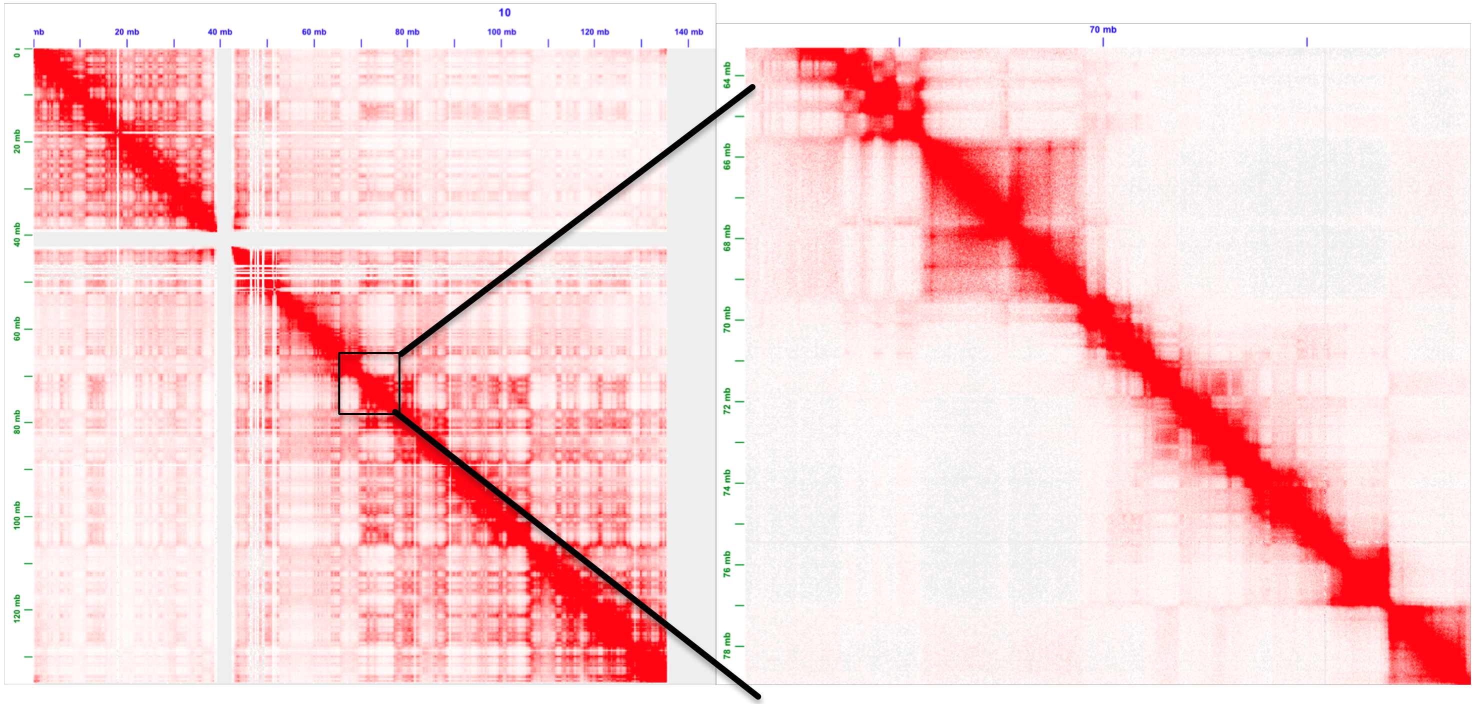
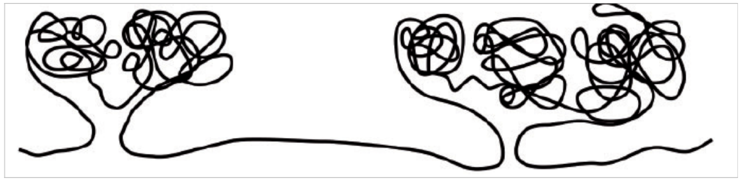

Tertiary structure
30nm chromatin Secondary structure
Nucleus with distinct territories
Nucleosome
DNA
10nm chromatin Primary Structure

image credit: Iyer et al. BMC Biophysics 2011

Topologically associating domains (TADs)

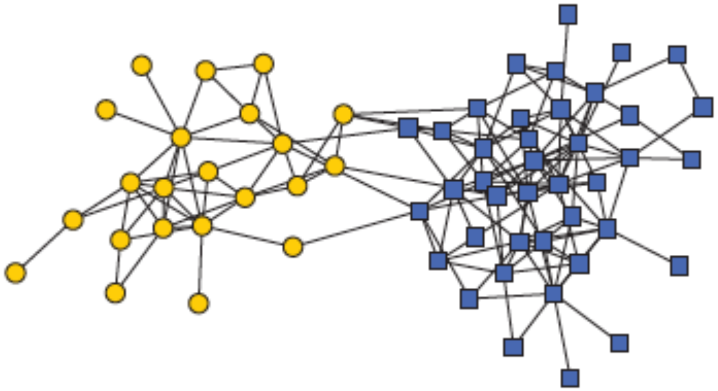


TADs have apparent hierarchical organization

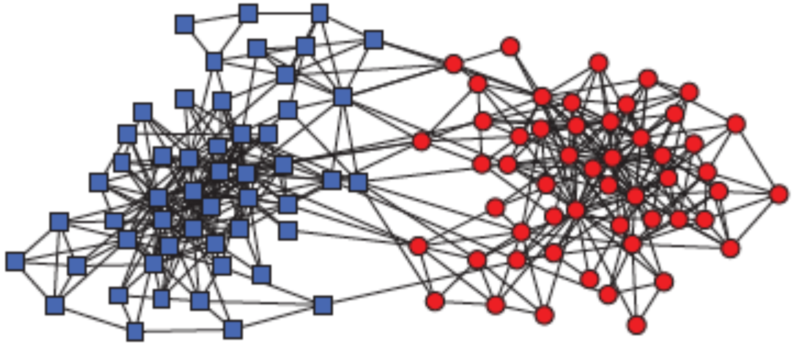


Modularity

Network modularity



Dolphin social network



Political books

Newman Phy. Rev. E 2013

degree of node i

adjacency matrix

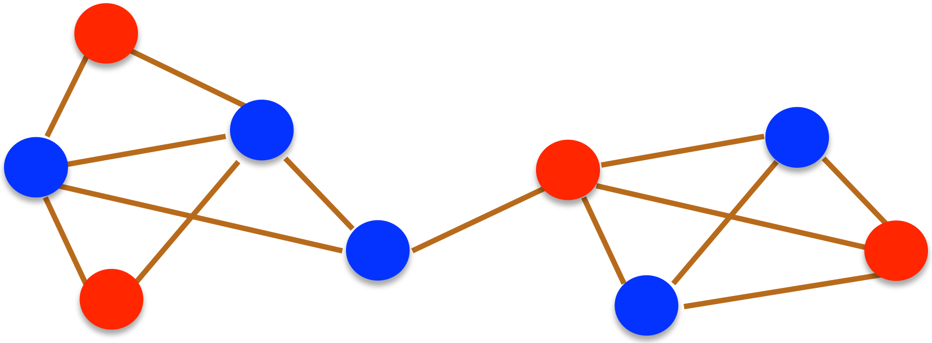
$$Q = \frac{1}{2m}\sum_{i,j}\left(W_{ij} - \boxed{\frac{k_i k_j}{2m}}\right)\delta_{\sigma_i \sigma_j}$$

whether or not
i, j are in the
same module

number of edges

expected number of
edges between i and j

Network modularity



$$Q \approx 0$$

degree of node i

adjacency matrix

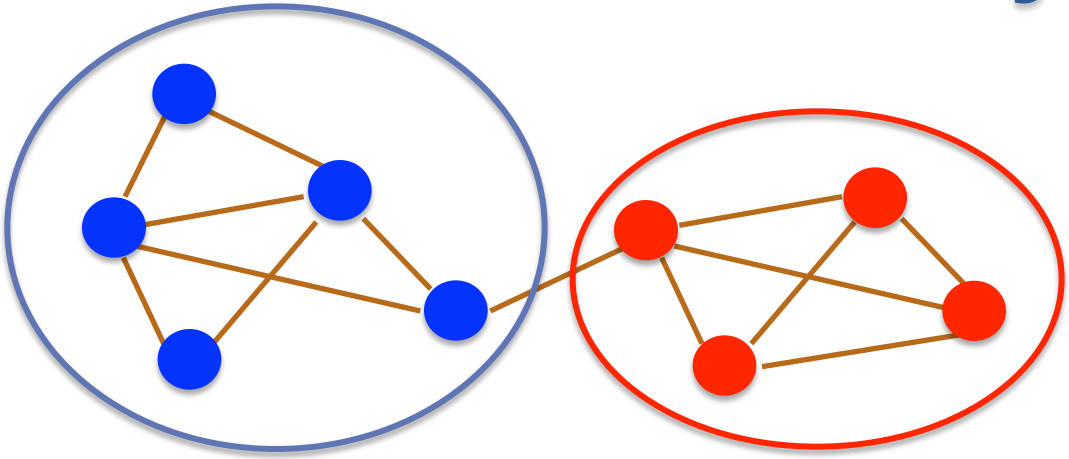
$$Q = \frac{1}{2m} \sum_{i,j} \left(W_{ij} - \boxed{\frac{k_i k_j}{2m}} \right) \delta_{\sigma_i \sigma_j}$$

whether or not i, j are in the same module

number of edges

expected number of edges between i and j

Network modularity



$$Q = Q_{max}$$

Optimization
problem
for sim.
annealing

degree of node i

adjacency matrix

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

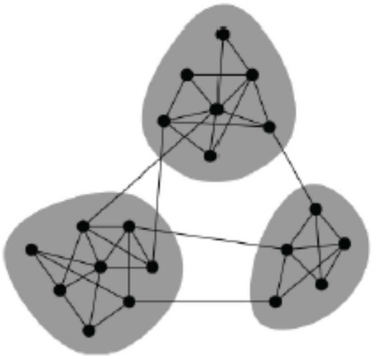
whether or not
i, j are in the same
module

number of edges

expected number of
edges between i and j

TAD Finding

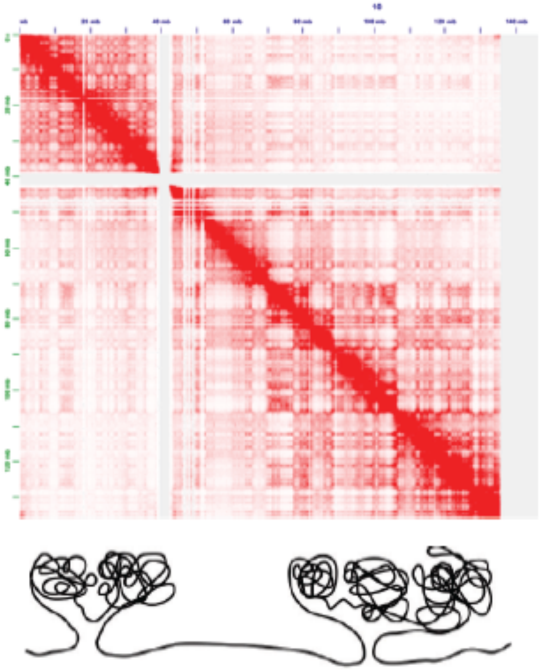
Identifying TADs in multiple resolutions



Modularity maximization

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

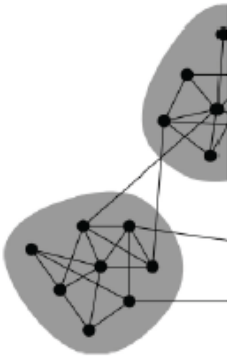
network	contact map
node	chromosome bin
edge	Hi-C contact
# of connections	coverage
module	domain



schematic adapted from ref. [2]

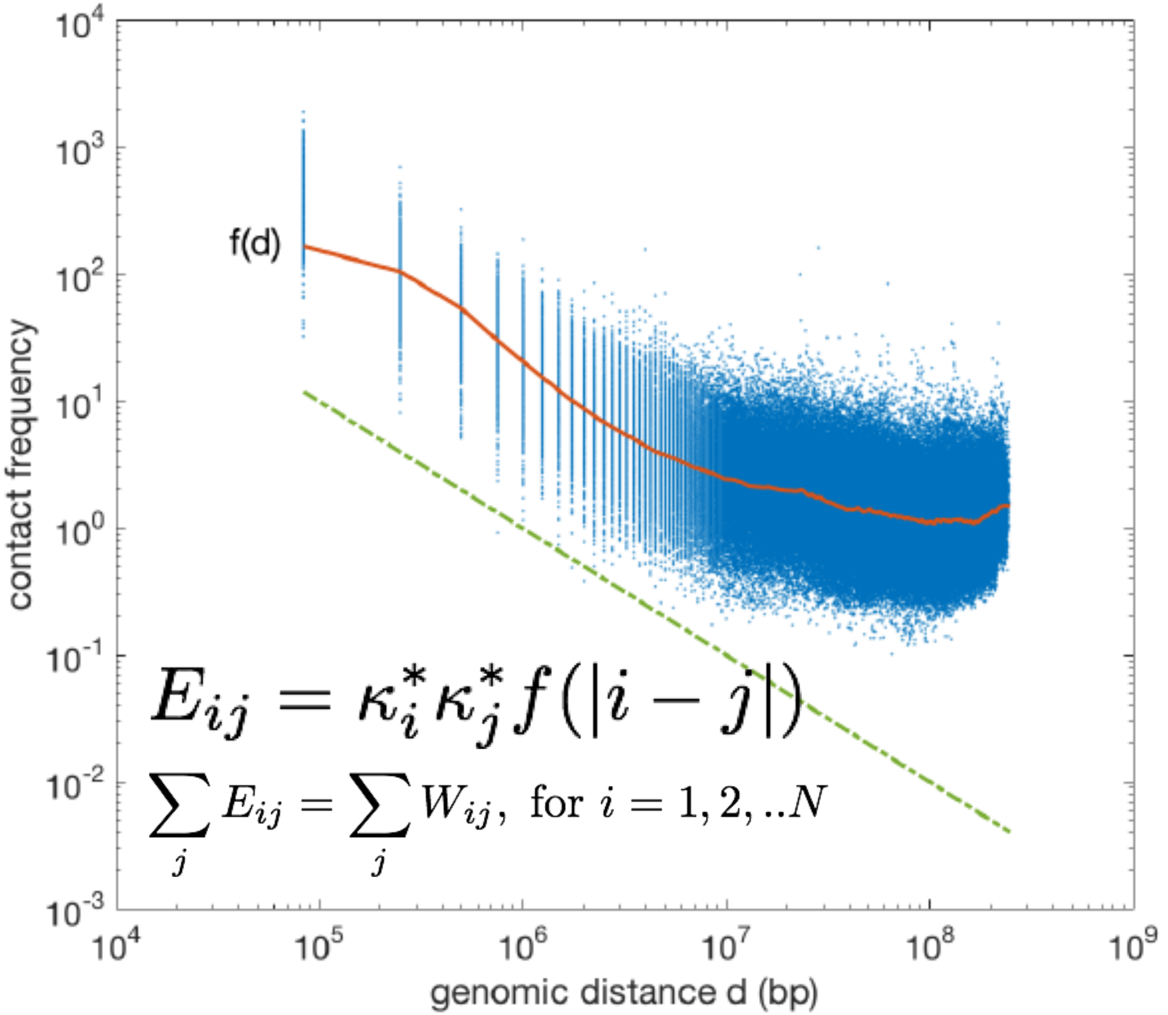
[Yan et al., *PLOS Comp. Bio.* (in revision, ‘17); bioRxiv 097345]

Identifying TADs in multiple resolutions



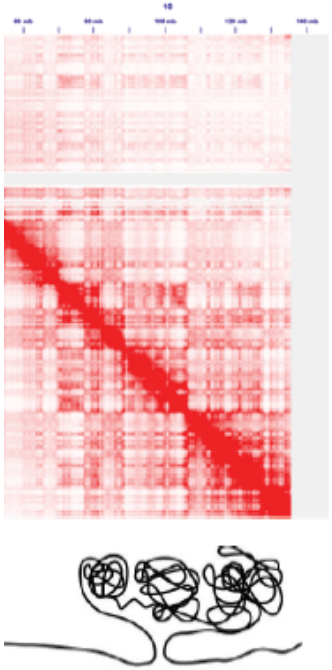
Modularity m

$$Q = \frac{1}{2m} \sum_{i,j}$$



$$E_{ij} = \kappa_i^* \kappa_j^* f(|i-j|)$$

$$\sum_j E_{ij} = \sum_j W_{ij}, \text{ for } i = 1, 2, ..N$$

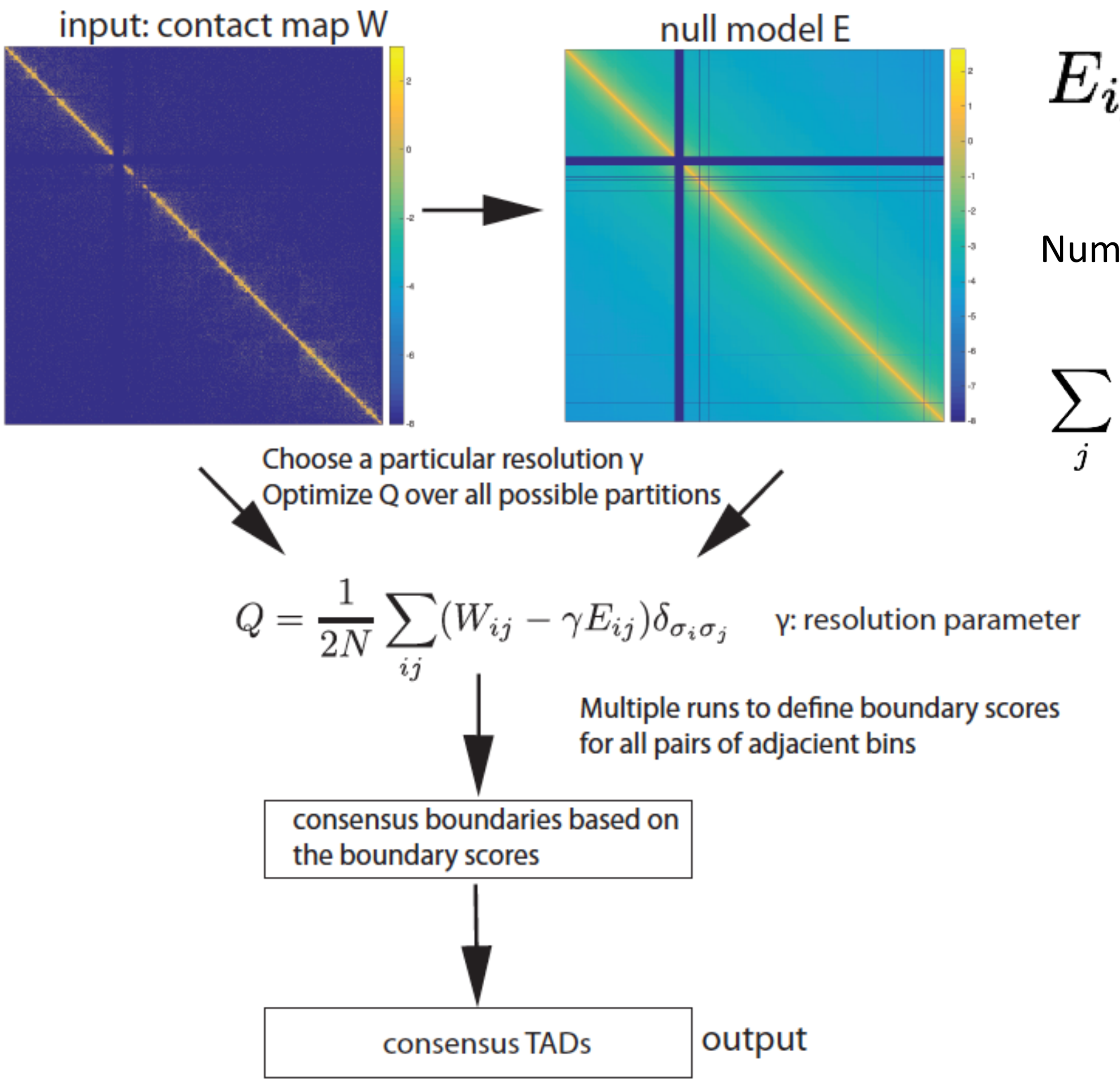


: adapted from ref. [2]

[Yan et al., *PLOS Comp. Bio.* (in revision, ‘17); bioRxiv 097345]

[Yan et al., *PLOS Comp. Bio.* (in revision, ‘17); bioRxiv 097345]

Identifying TADs in multiple resolutions



Choose a particular resolution γ
Optimize Q over all possible partitions

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

γ: resolution parameter

Multiple runs to define boundary scores for all pairs of adjacent bins

consensus boundaries based on the boundary scores

consensus TADs

output

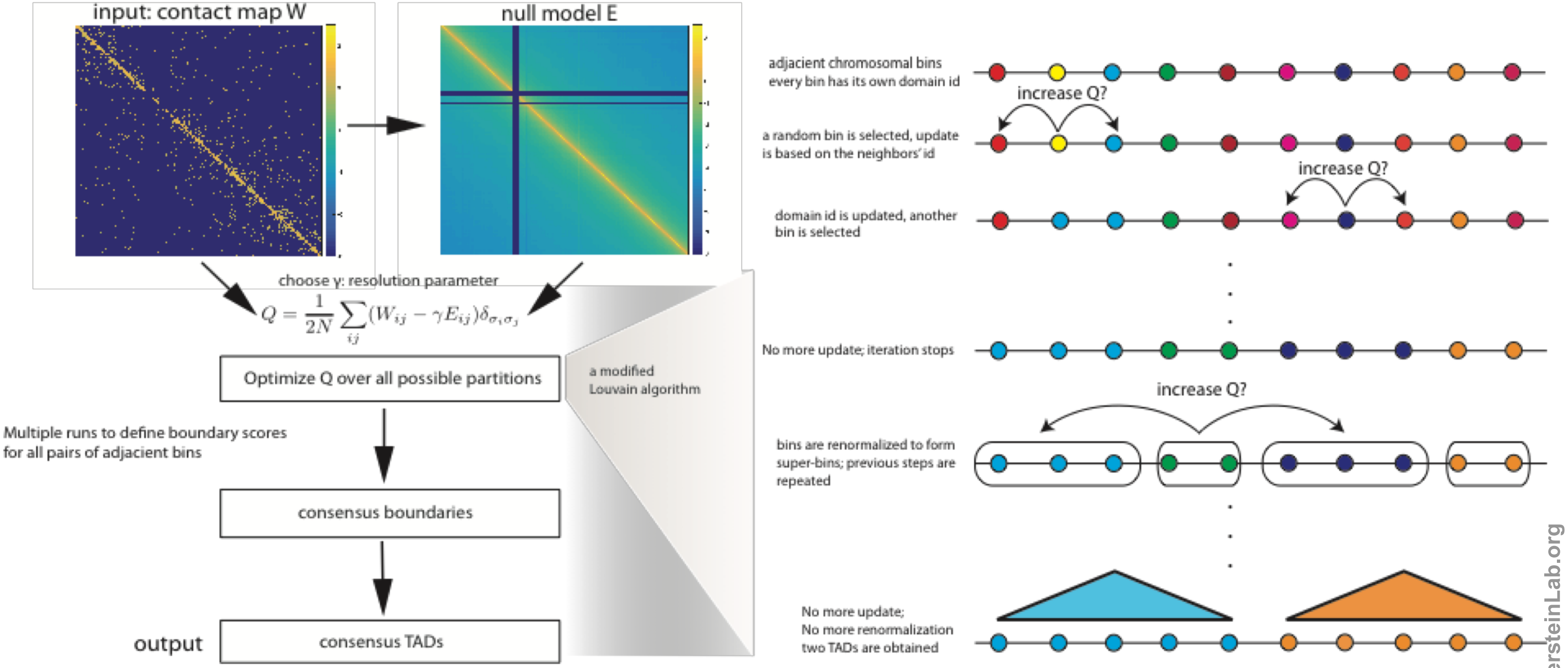
$$E_{ij} = \kappa_i^* \kappa_j^* f(|i-j|)$$

Numerically solve for κ_i^* in equations

$$\sum_j E_{ij} = \sum_j W_{ij}, \text{ for } i = 1, 2, ..N$$

[Yan et al., *PLOS Comp. Bio.* (in revision, ‘17); bioRxiv 097345]

Identifying TADs in multiple resolutions



input: contact map W

null model E

choose γ: resolution parameter

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

Optimize Q over all possible partitions

a modified Louvain algorithm

Multiple runs to define boundary scores for all pairs of adjacent bins

consensus boundaries

output

consensus TADs

adjacent chromosomal bins every bin has its own domain id

increase Q?

a random bin is selected, update is based on the neighbors' id

increase Q?

domain id is updated, another bin is selected

No more update; iteration stops

increase Q?

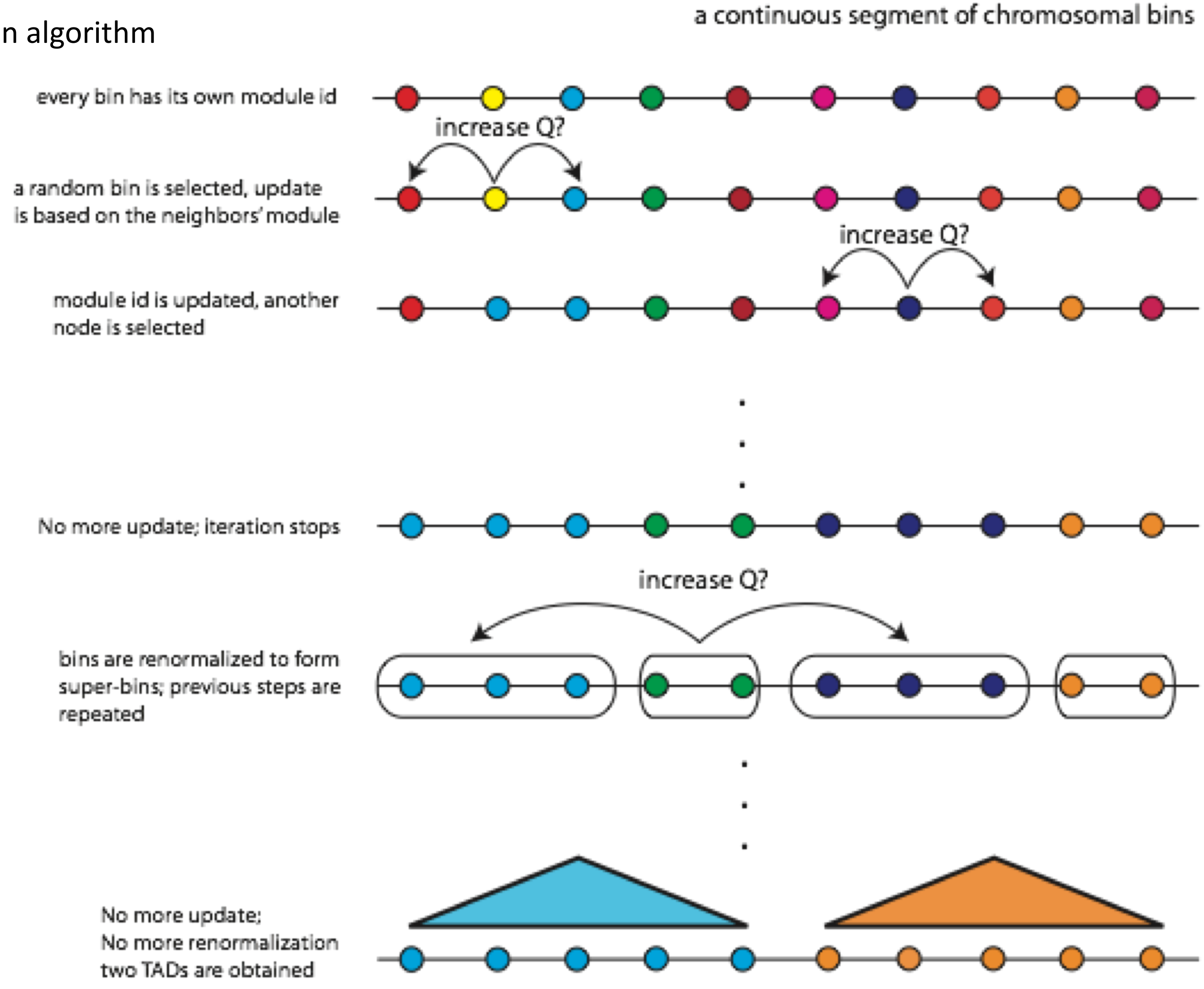
bins are renormalized to form super-bins; previous steps are repeated

No more update; No more renormalization two TADs are obtained

[Yan et al., *PLOS Comp. Bio.* (in revision, ‘17); bioRxiv 097345]

Identifying TADs in multiple resolutions

a modified Louvain algorithm



a continuous segment of chromosomal bins

every bin has its own module id

a random bin is selected, update is based on the neighbors' module

module id is updated, another node is selected

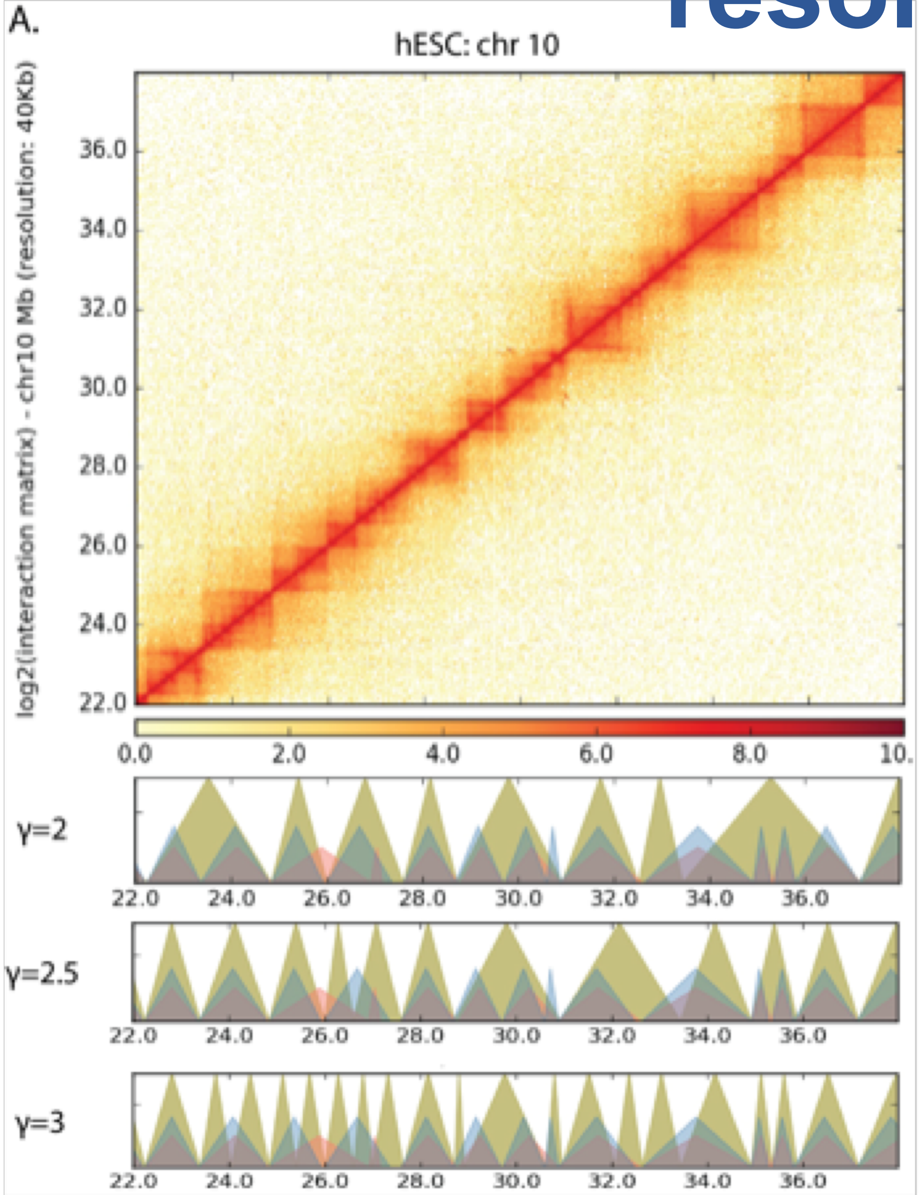
No more update; iteration stops

bins are renormalized to form super-bins; previous steps are repeated

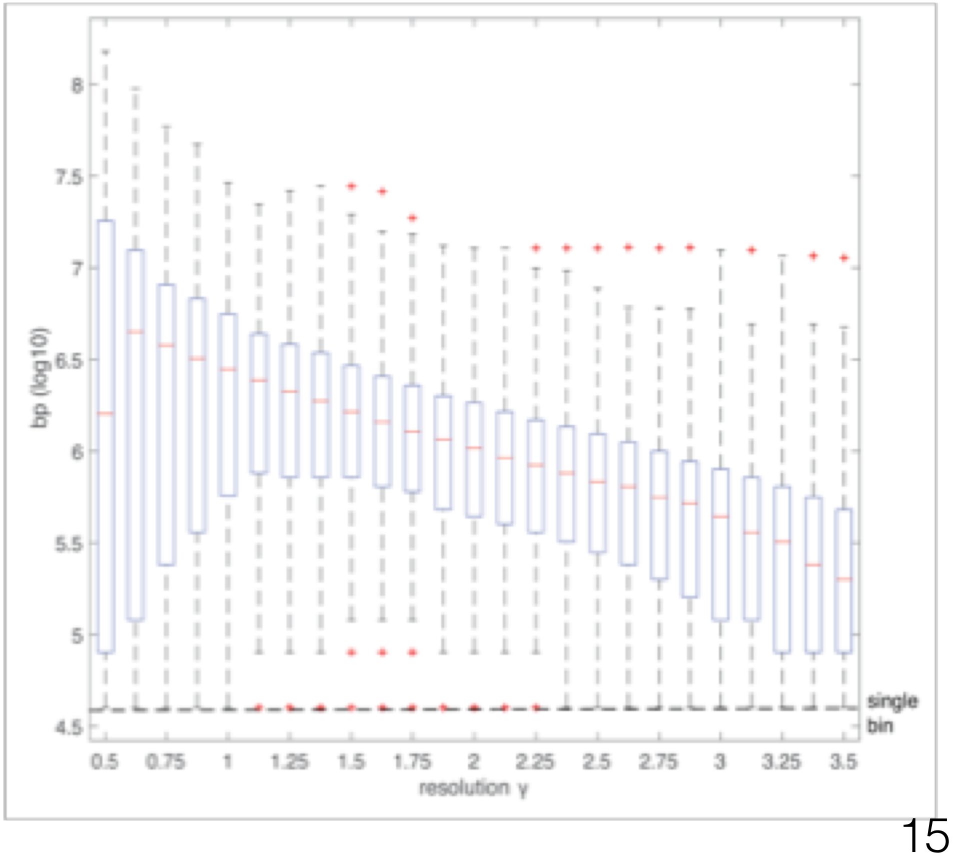
No more update; No more renormalization two TADs are obtained

Identifying TADs in multiple resolutions

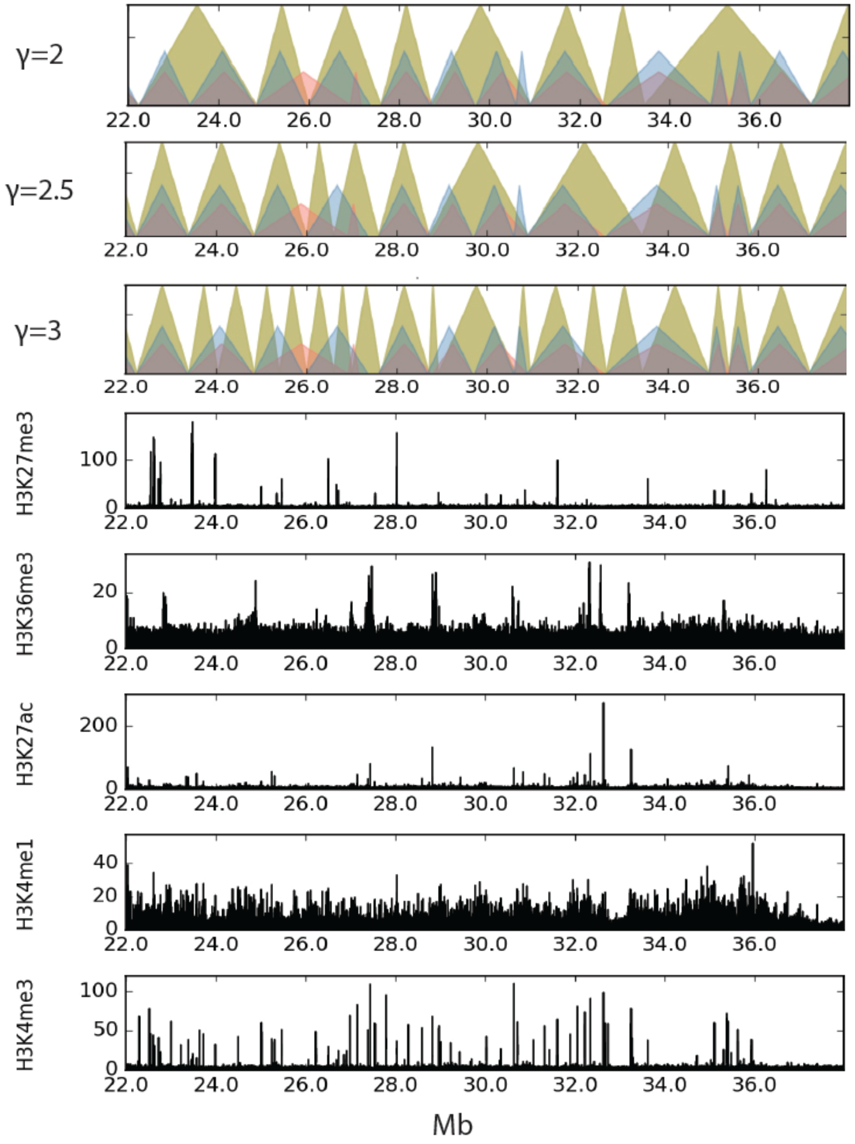
[Yan et al., *PLOS Comp. Bio.* (in revision, ‘17); bioRxiv 097345]

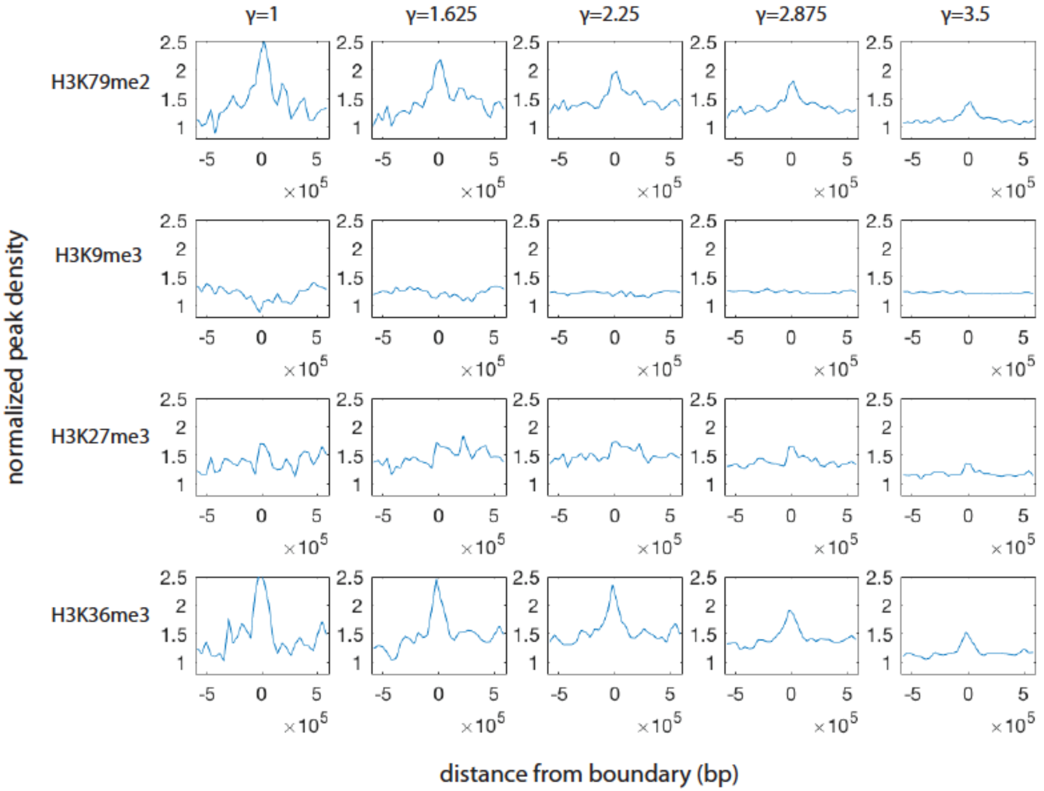


smaller TADs but are detected as the resolution increases



Enrichment of histone features at different resolution



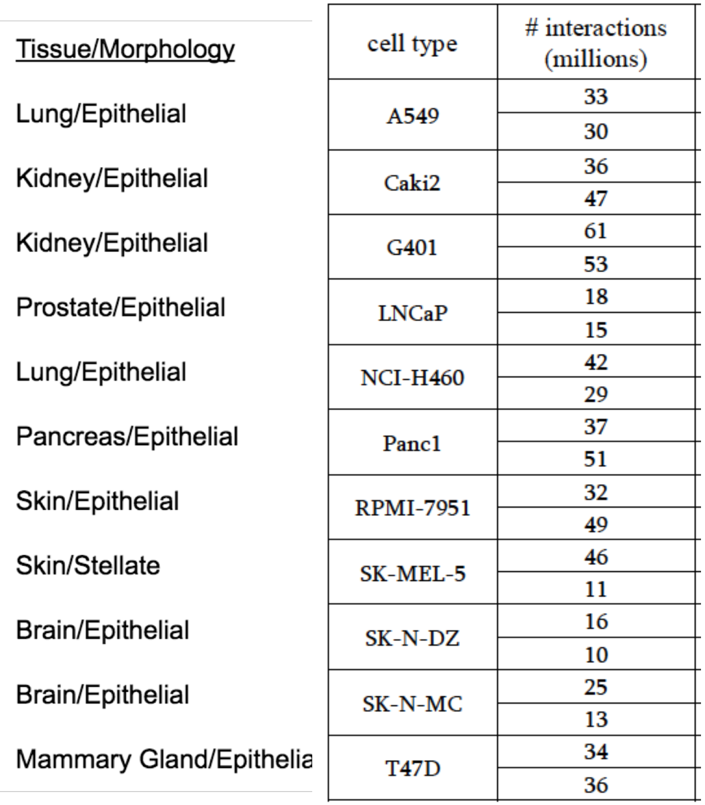


[Yan et al., *PLOS Comp. Bio.* (in revision, '17); bioRxiv 097345]

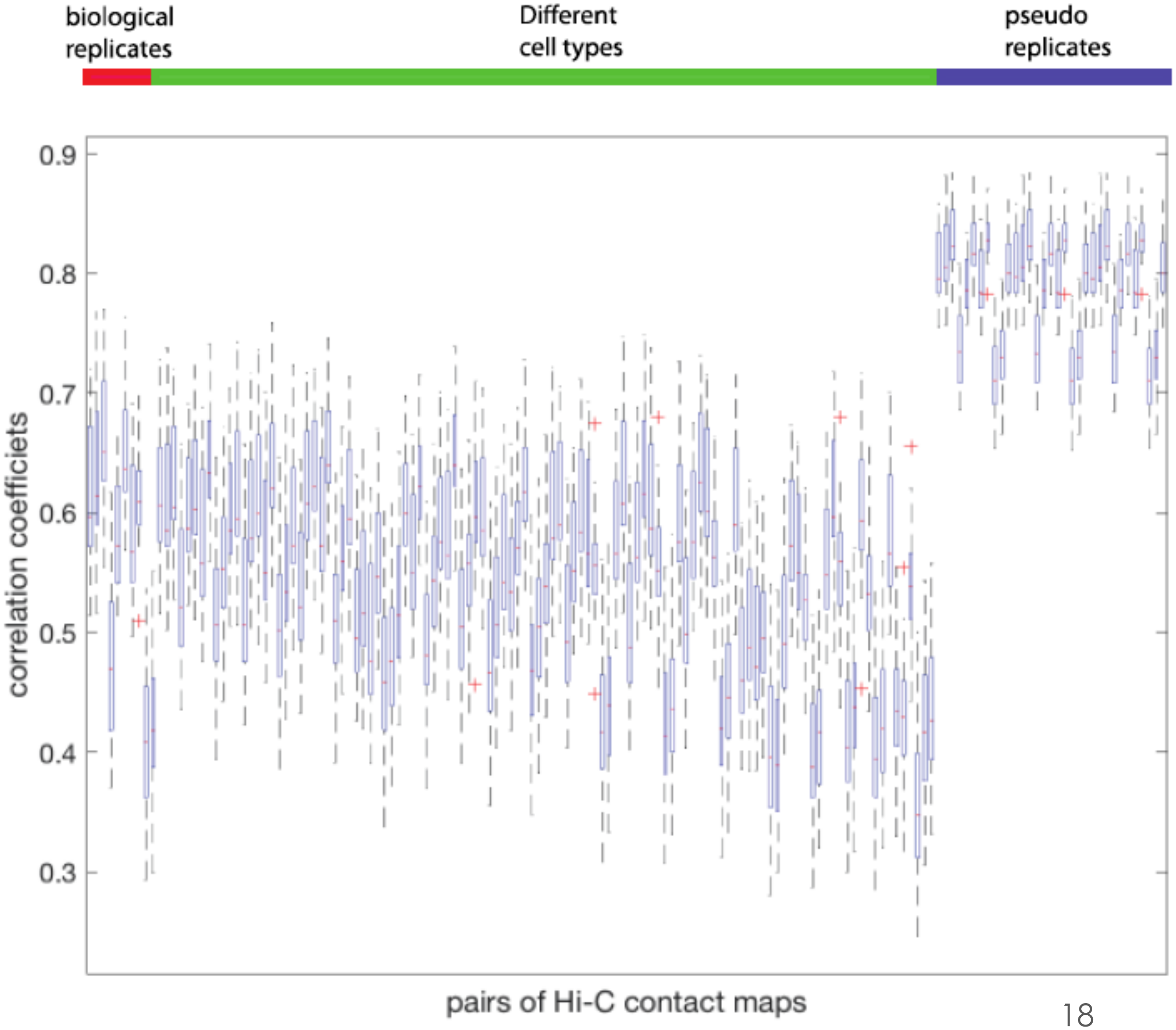
Using Matrix Decomposition for Hi-C Contact Matrices

Quantifying reproducibility of Hi-C data

ENCODE Hi-C data



Tissue/Morphology	cell type	# interactions (millions)
Lung/Epithelial	A549	33
		30
Kidney/Epithelial	Caki2	36
		47
Kidney/Epithelial	G401	61
		53
Prostate/Epithelial	LNCaP	18
		15
Lung/Epithelial	NCI-H460	42
		29
Pancreas/Epithelial	Panc1	37
		51
Skin/Epithelial	RPMI-7951	32
		49
Skin/Stellate	SK-MEL-5	46
		11
Brain/Epithelial	SK-N-DZ	16
		10
Brain/Epithelial	SK-N-MC	25
		13
Mammary Gland/Epithelia	T47D	34
		36



[Yan et al., *Bioinformatics* ('17)]

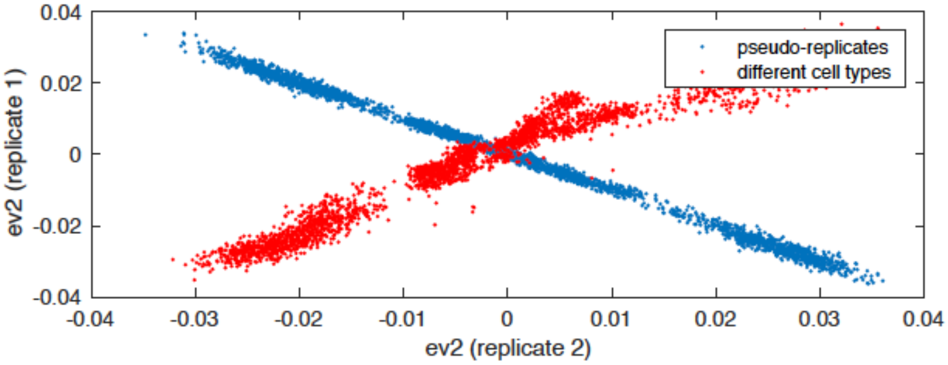
Quantifying reproducibility of Hi-C data

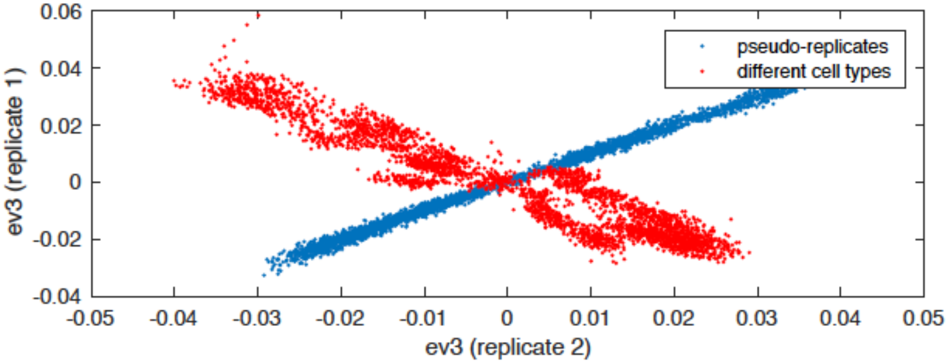
Is there a better way to decompose the contact map W (matrix)?

- Spectral clustering commonly used in image processing
- Transform W into the Laplacian matrix

$$\mathcal{L} = I - D^{-1/2} W D^{-1/2}, D_{ii} = \sum_j W_{ij}$$

- Decomposed into eigenvectors, and consider only the leading ones (dimension reduction)
- Distance between the corresponding vectors

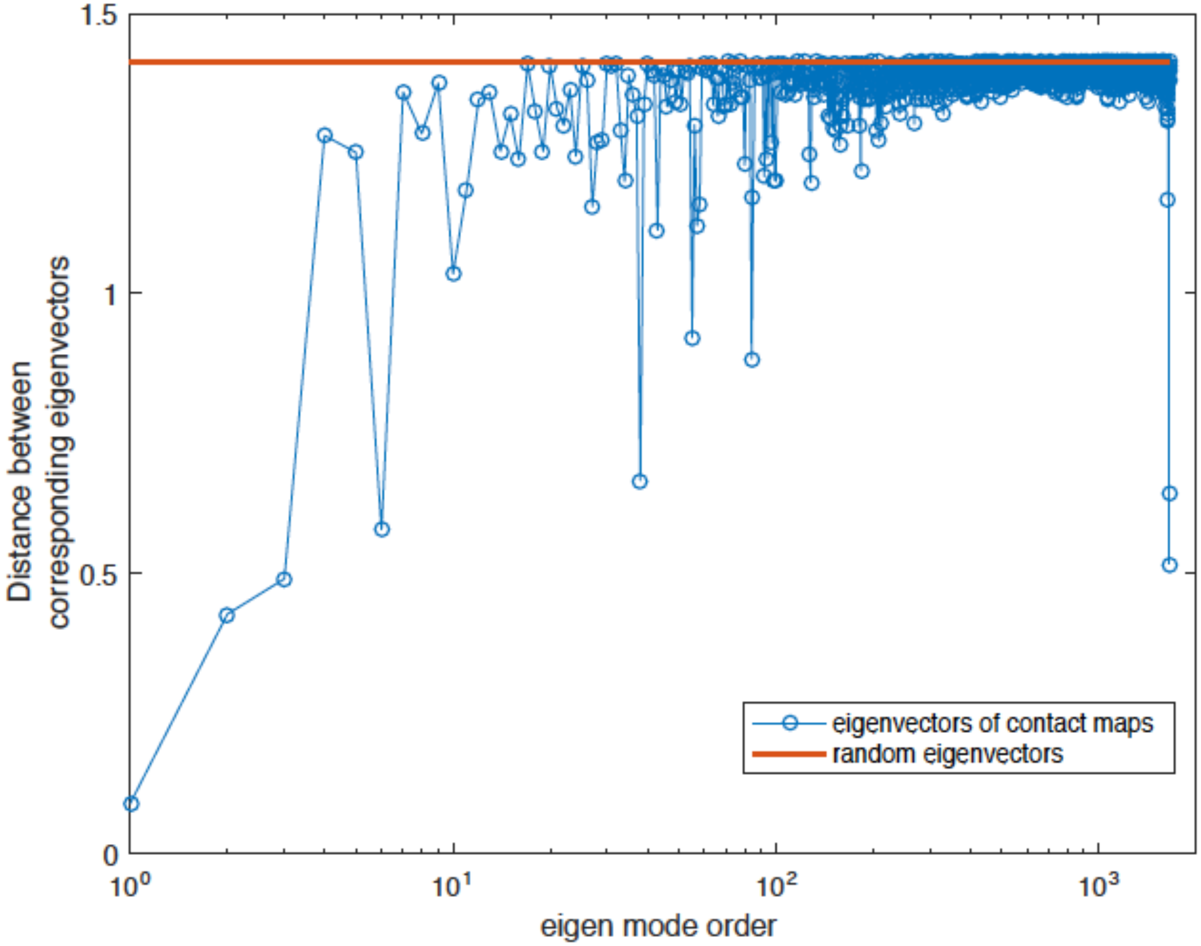




Yan KK et al. Bioinformatics 2017

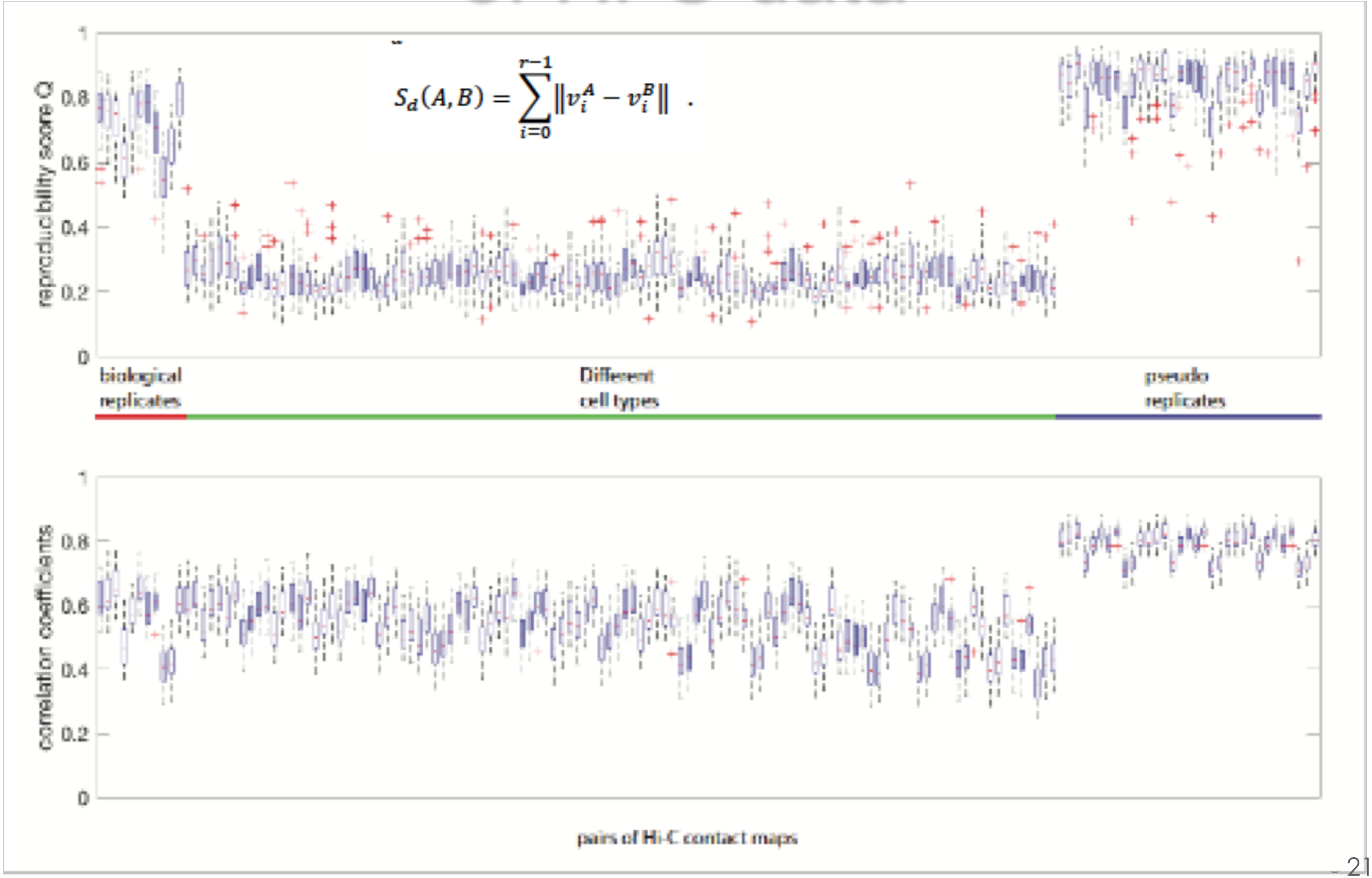
Quantifying reproducibility of Hi-C data

How many eigenvectors should be used?



Yan KK et al. Bioinformatics 2017

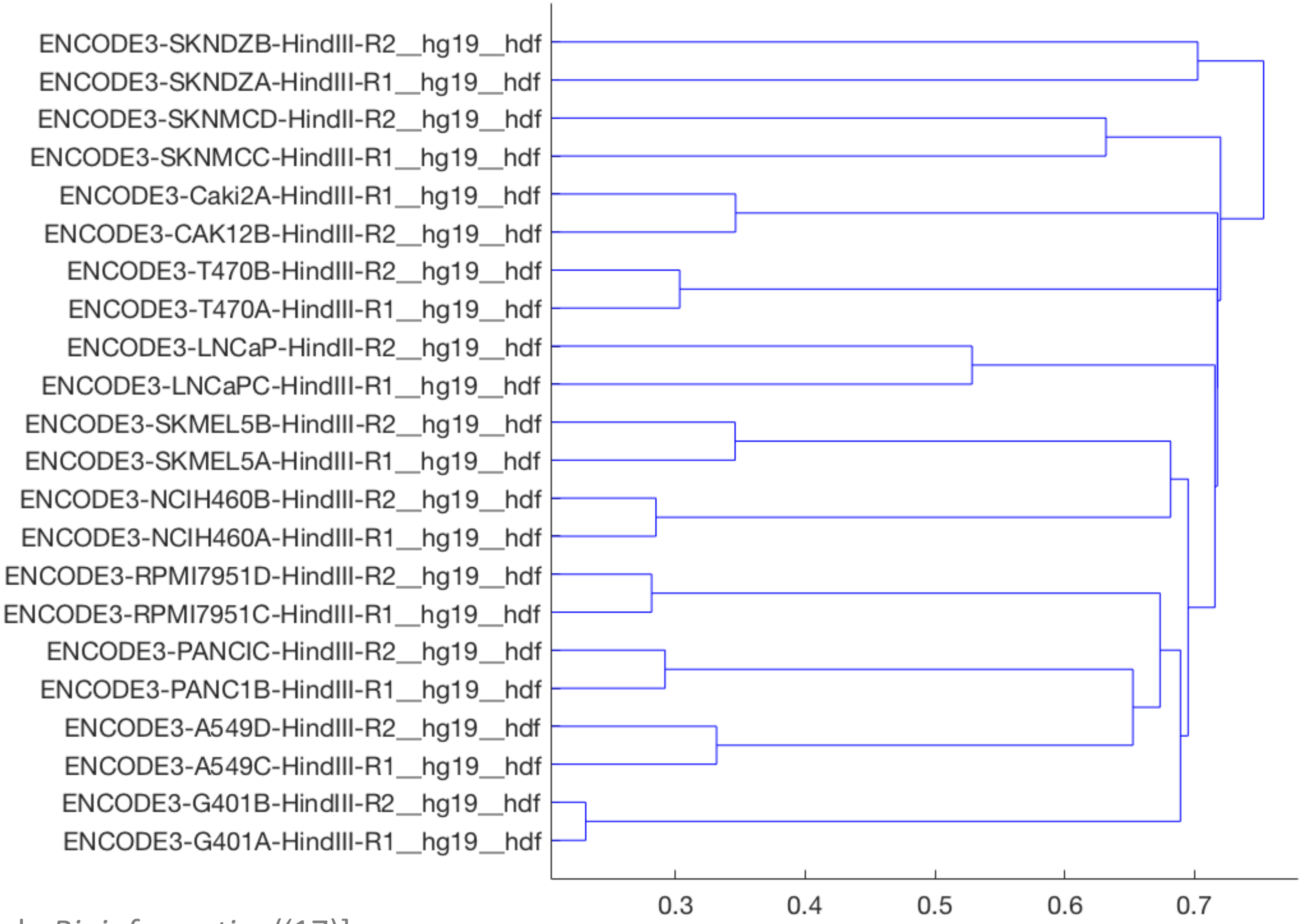
Quantifying reproducibility of Hi-C data



$$S_d(A,B) = \sum_{i=0}^{r-1} \left\| v_i^A - v_i^B \right\| \ .$$

Yan KK et al. Bioinformatics 2017

A distance measure between two contact maps



[Yan et al., *Bioinformatics* ('17)]