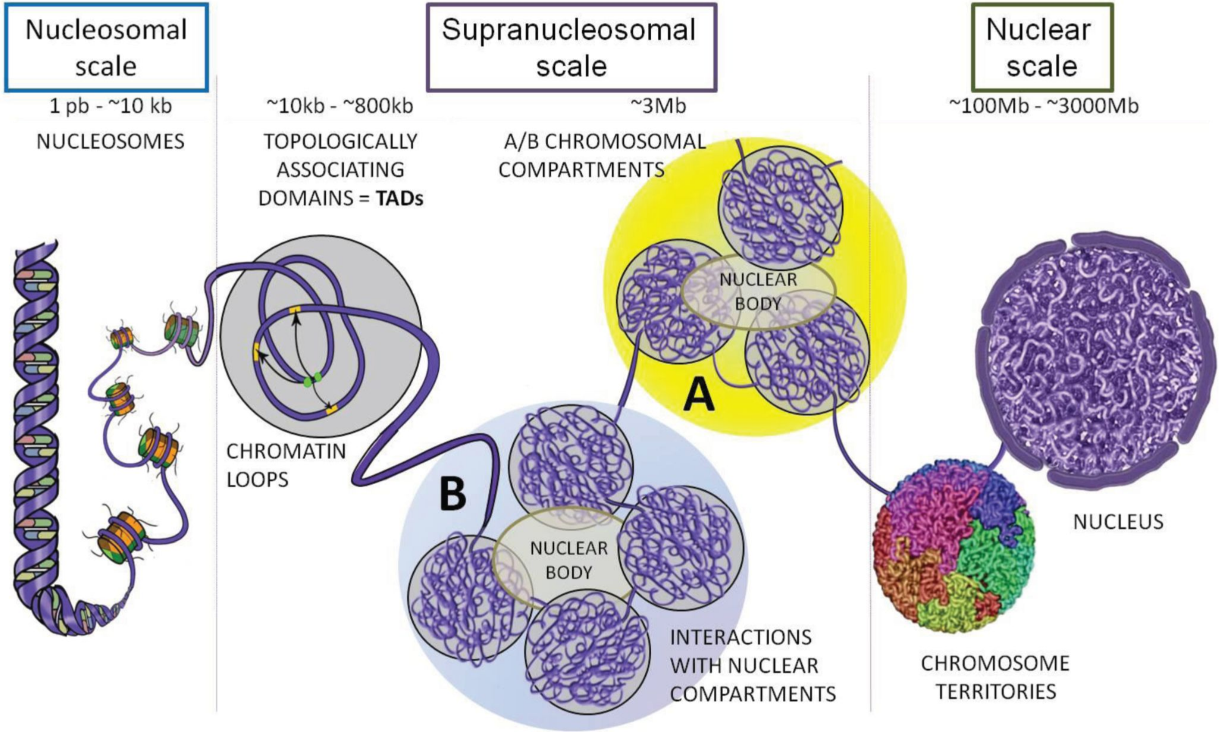
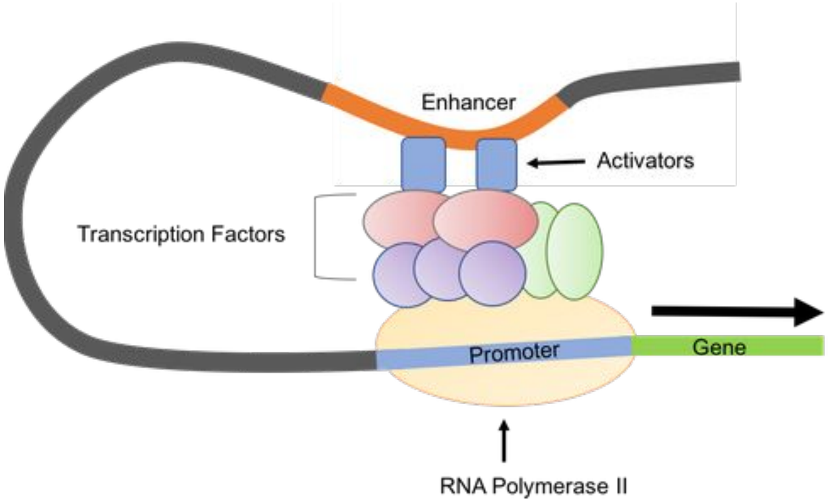
Analysis of Hi-C data using numerical linear algebra approaches

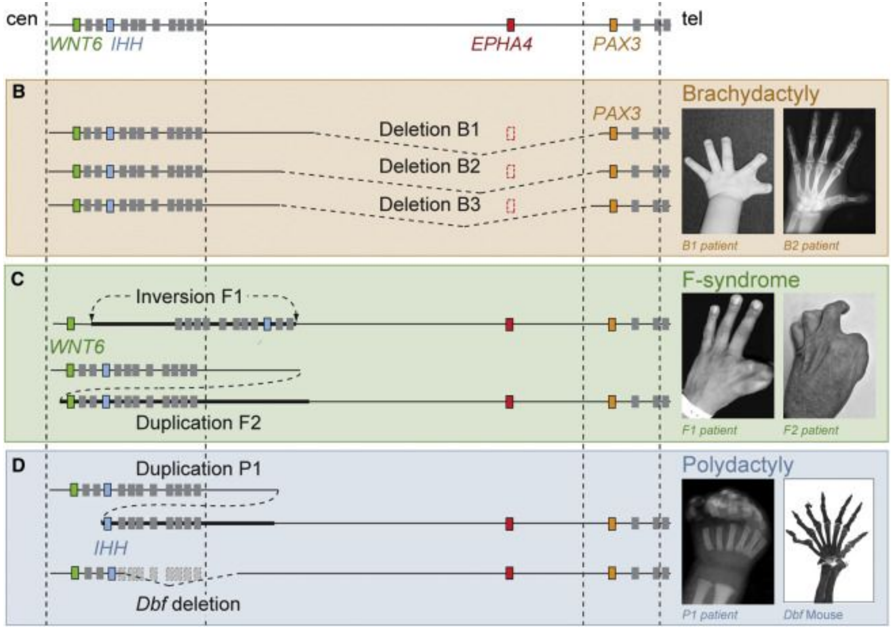
Kirill Ulianov
Marina Morozova
Maksim Grigoryan
Shamil Magomedov

A few words about DNA…

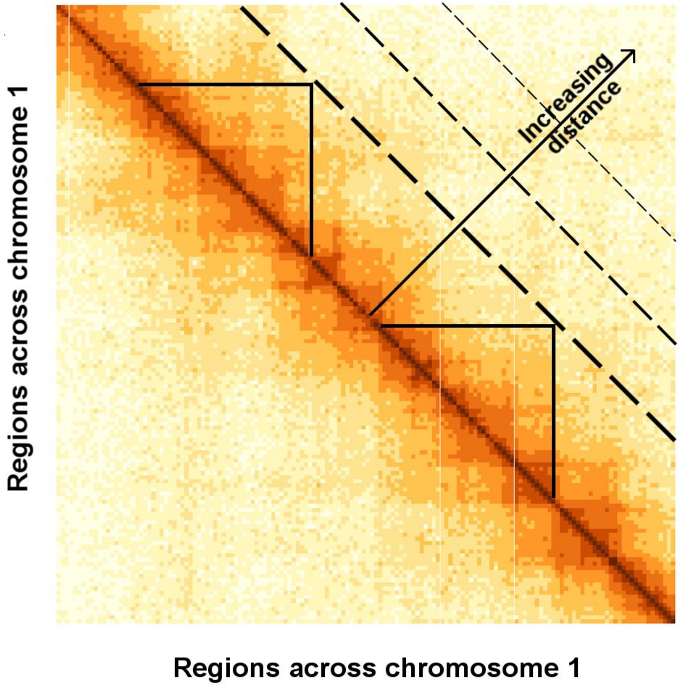


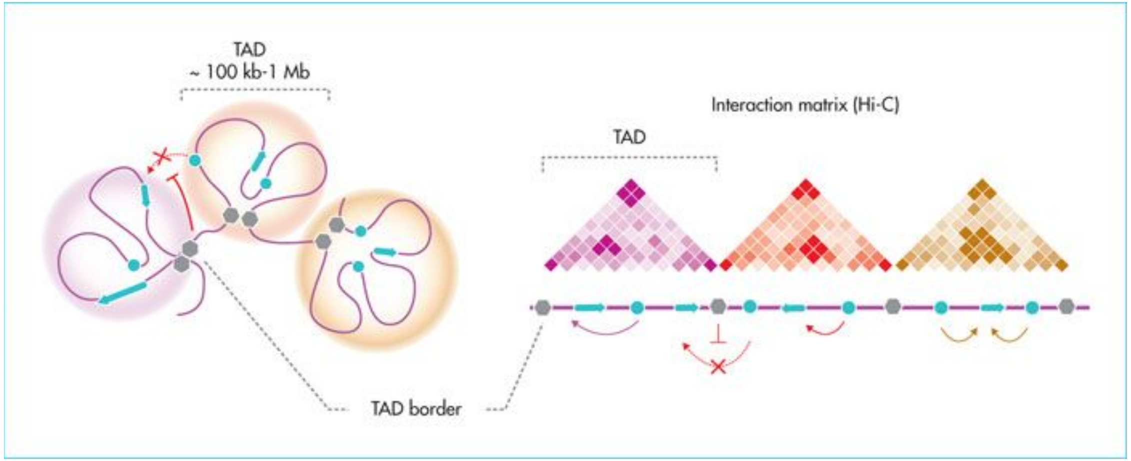
Why does it matter?





What is a Hi-C matrix?





Example of a related task

One is common purpose in Hi-C analysis is detecting of the large conglomerates of active and repressed genome parts - compartments.

The conventional approach (Lieberman-Aiden, 2009) utilises only the first eigenvector of preprocess Hi-C matrix as the main predictor of DNA compartments distribution

Some methods use all eigenvectors to restore compartment, e.g. Gaussian Network Model (GNM).

Our **goal** was to use both approaches for Hi-C data and compare the results. Additionally, we were validated by gene activity patterns inferred from independent experiments

Data

For the project we selected two Hi-C matrix from a public database derived for human and yeast cell cultures. From each map we chose only one chromosome for the analysis.

The human 2nd chromosome was splitted on 4844 equally sized bins.

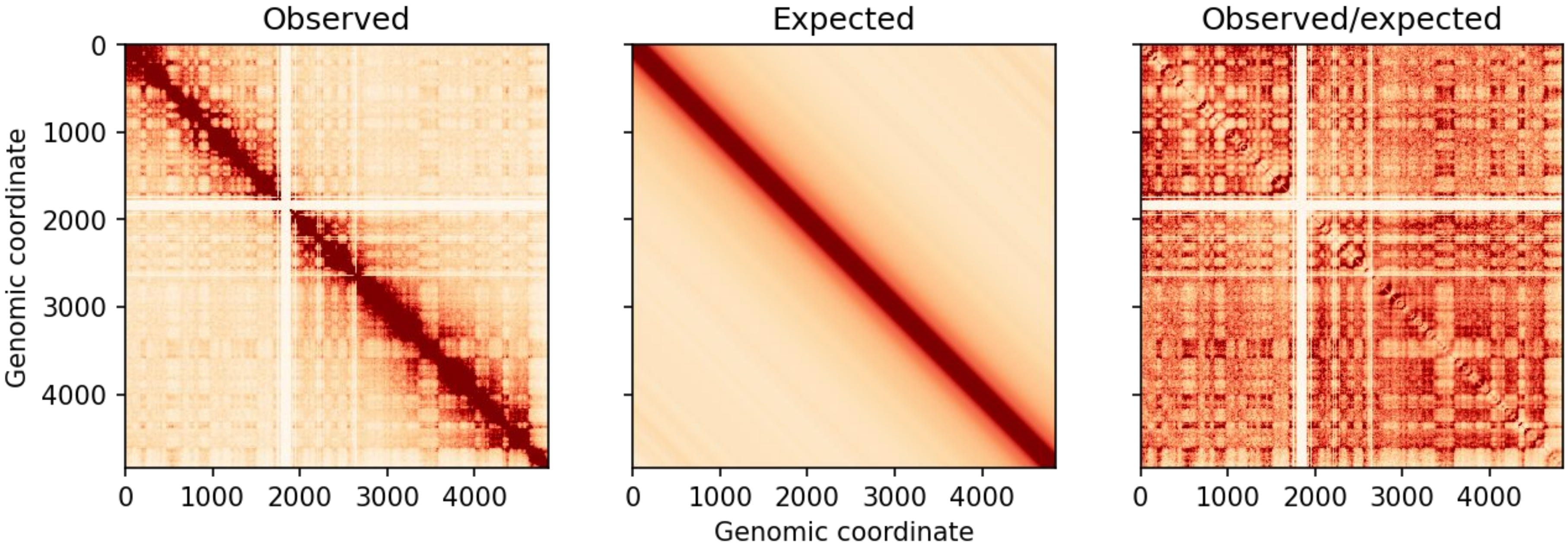
The yeast 5th chromosome was splitted on 722 equally sized bins.

Eigendecomposition

First eigenvector and properties of Hi-C matrix

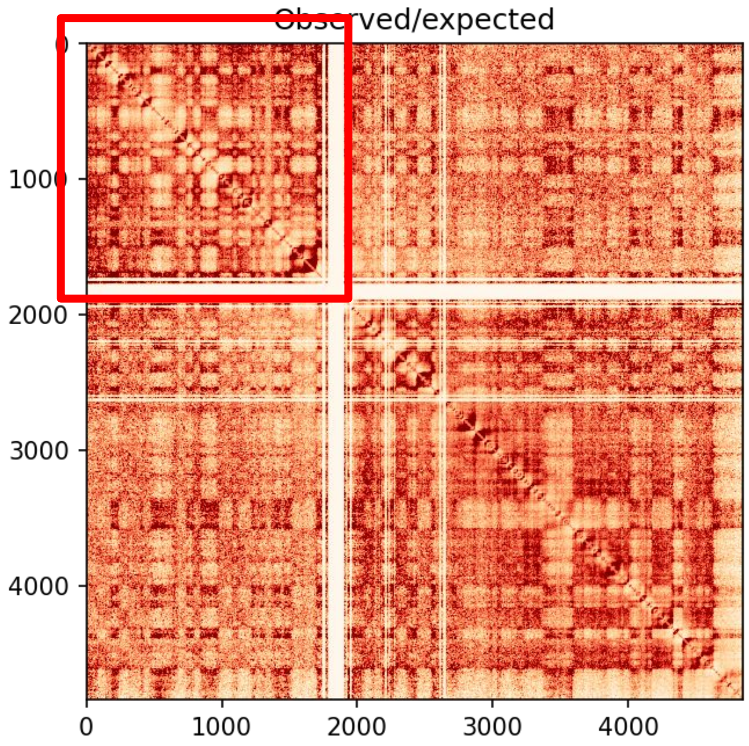
1) Observed/Expected values
2) Eigendecomposition

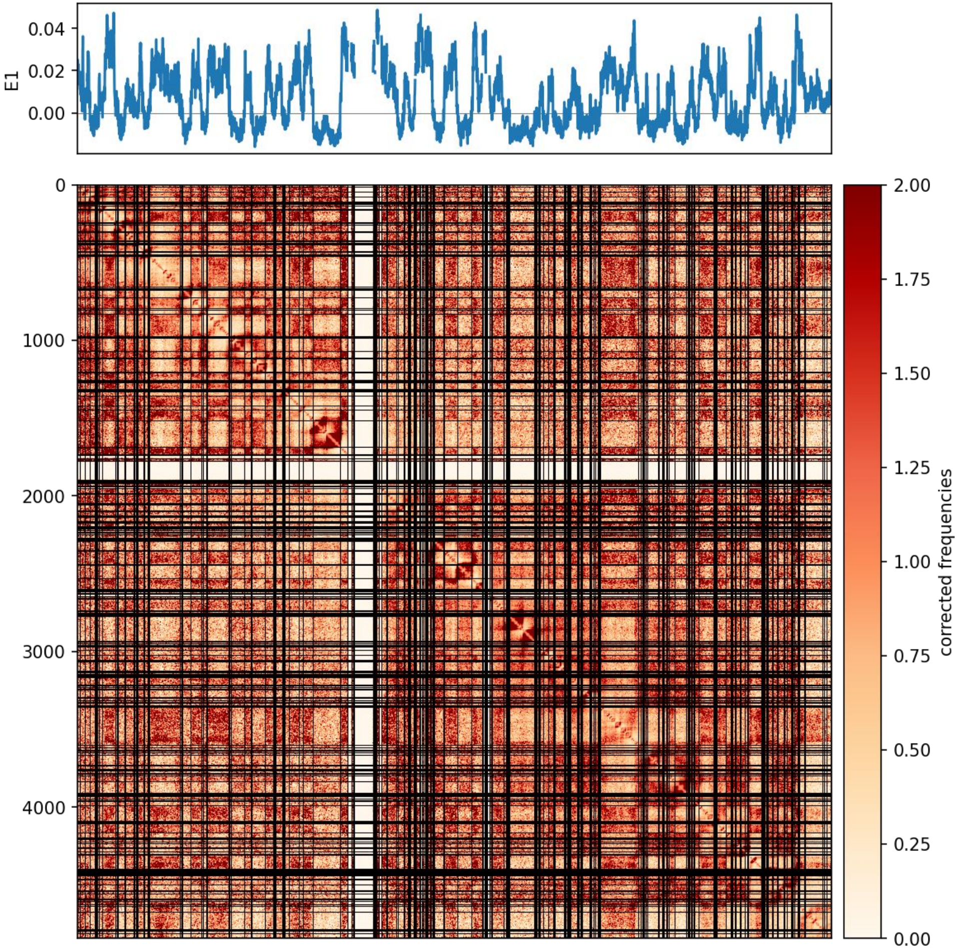
Human 2nd chromosome



Averaged diagonals of Observed
(Toeplitz matrix)

First eigenvector (E1)

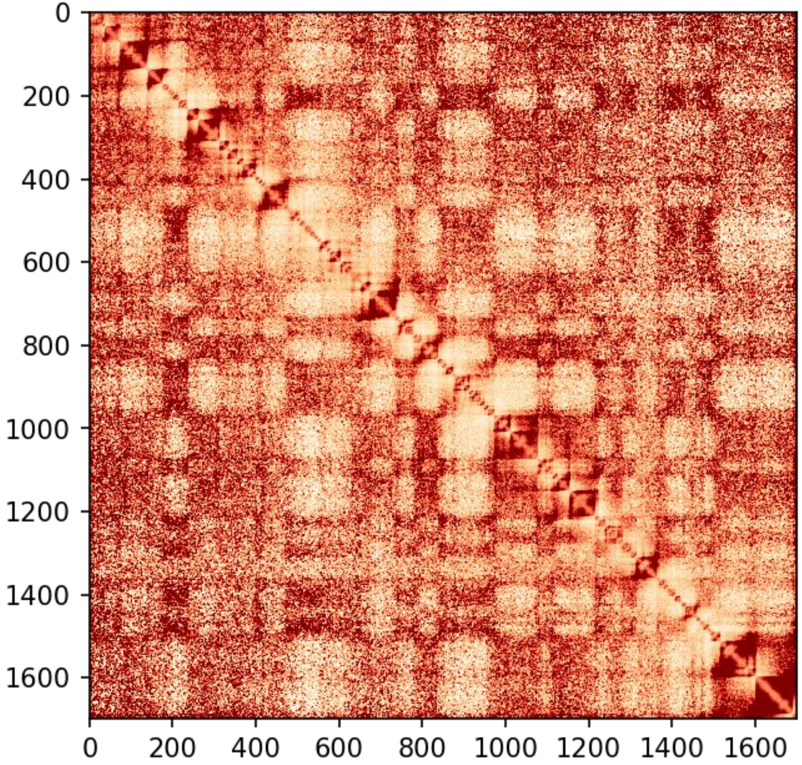


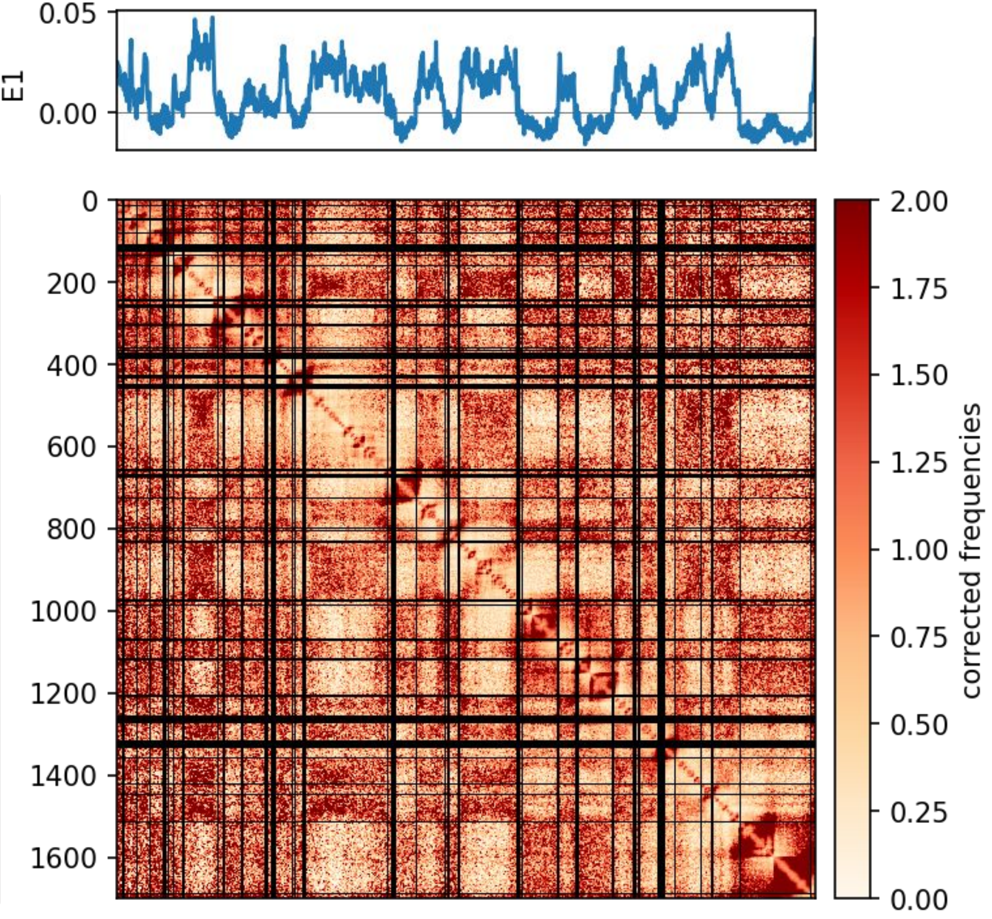


Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., … Mirny, L. A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nature Methods, 9(10), 999–1003. doi:10.1038/nmeth.2148 (https://doi.org/10.1038/nmeth.2148)

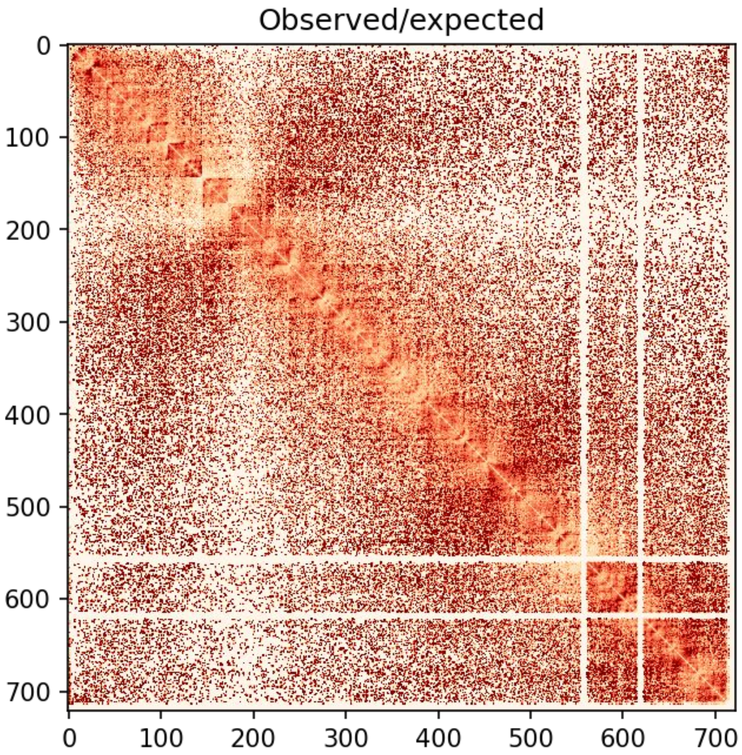
Closer look

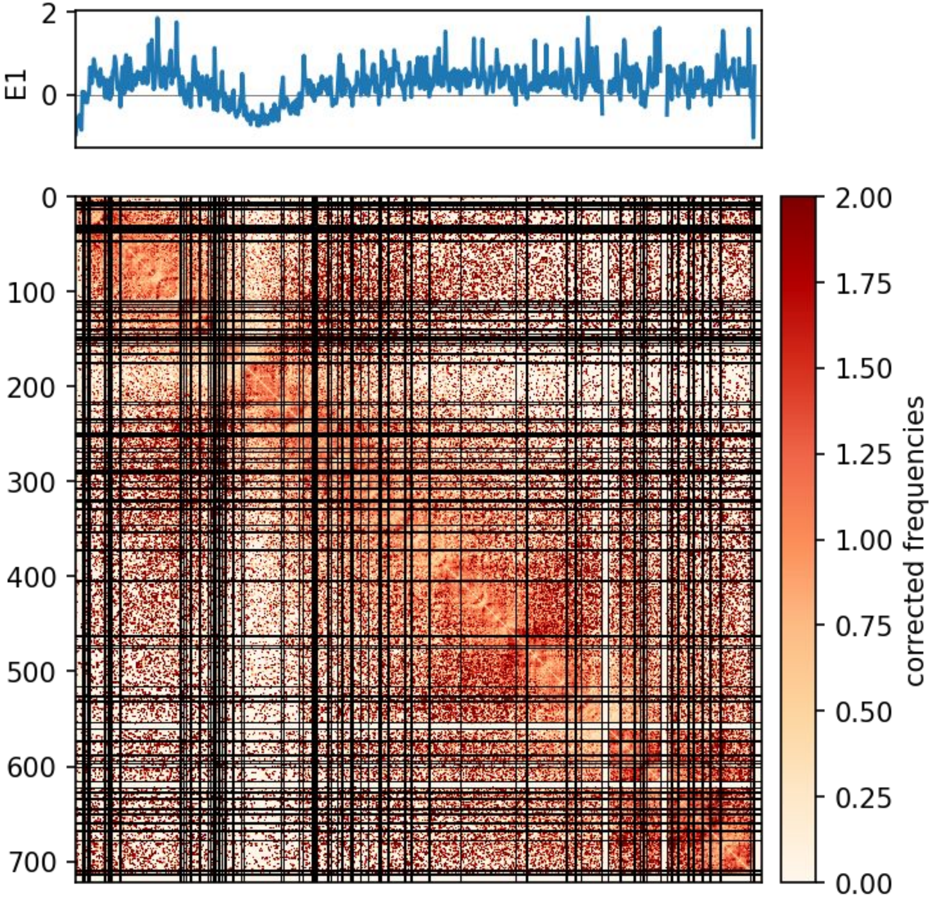
Observed/expected;
a section of the matrix





Yeast 5th chromosome



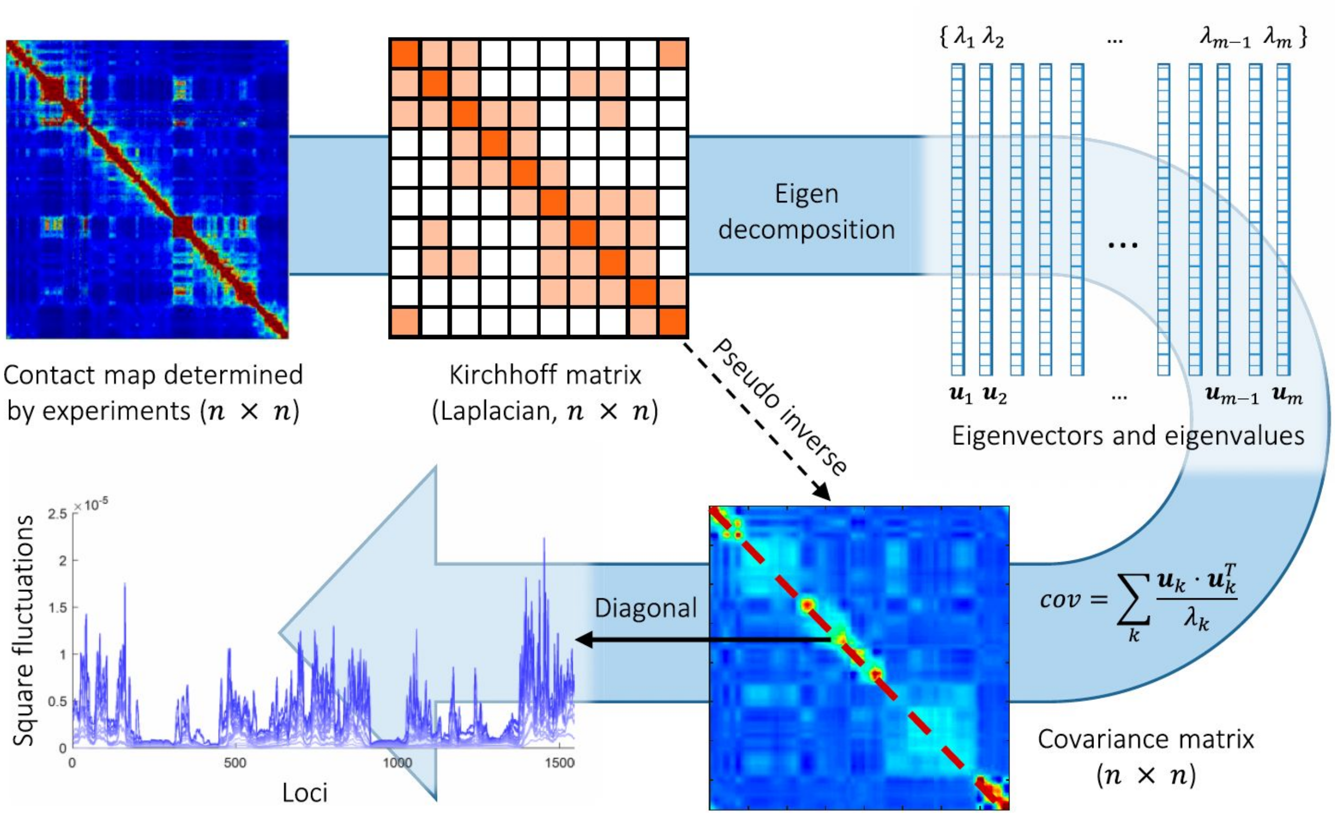


Gaussian network model

Eigendecomposition
with extra steps

1) Kirchhoff matrix
2) Covariance matrix
3) Mean-square fluctuation

Gaussian network model



Contact map determined by experiments ($n \times n$)

Kirchhoff matrix (Laplacian, $n \times n$)

Eigen decomposition

$\{\lambda_1\ \lambda_2\ \dots\ \lambda_{m-1}\ \lambda_m\}$

$\boldsymbol{u}_1\ \boldsymbol{u}_2\ \dots\ \boldsymbol{u}_{m-1}\ \boldsymbol{u}_m$

Eigenvectors and eigenvalues

Pseudo inverse

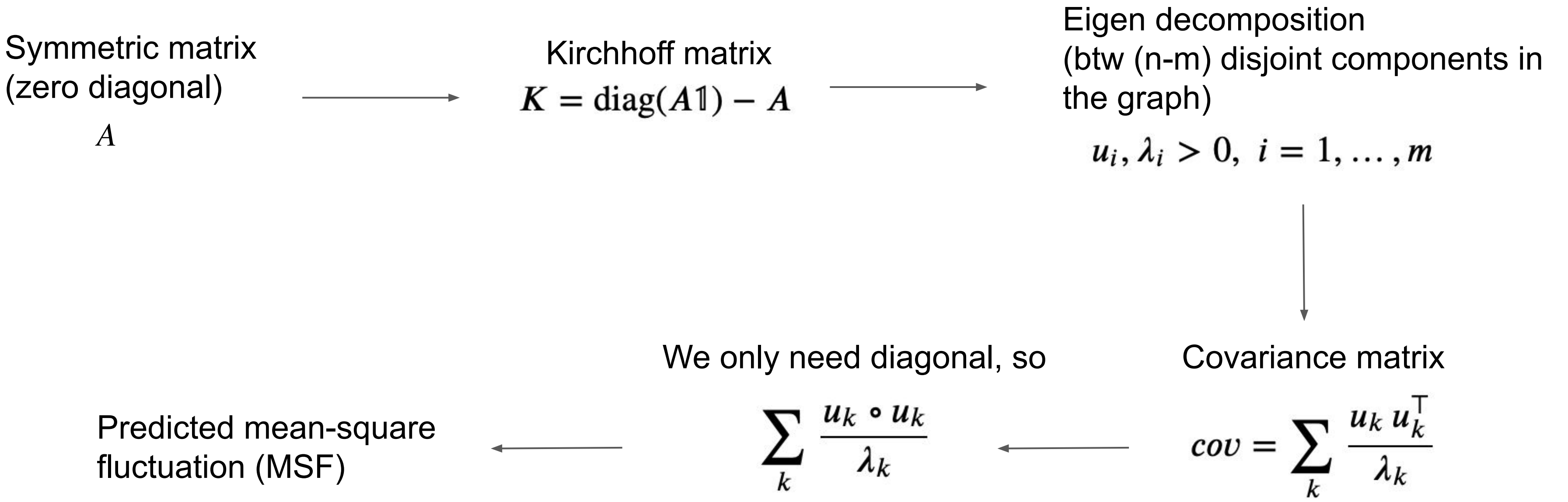
$$cov = \sum_k \frac{\boldsymbol{u}_k \cdot \boldsymbol{u}_k^T}{\lambda_k}$$

Covariance matrix ($n \times n$)

Diagonal

Chromosomal dynamics predicted by an elastic network model explains genome-wide accessibility and long-range couplings. Sauerwald, Natalie et al in Nucleic Acids Research (2017) pp. 3663-3673. doi: 10.1093/nar/gkx172 (https://doi.org/10.1093/nar/gkx172)

Algorithm



Symmetric matrix (zero diagonal)

A

Kirchhoff matrix

$$K = \mathrm{diag}(A\mathbb{1}) - A$$

Eigen decomposition (btw (n-m) disjoint components in the graph)

$$u_i, \lambda_i > 0,\ i = 1, \dots, m$$

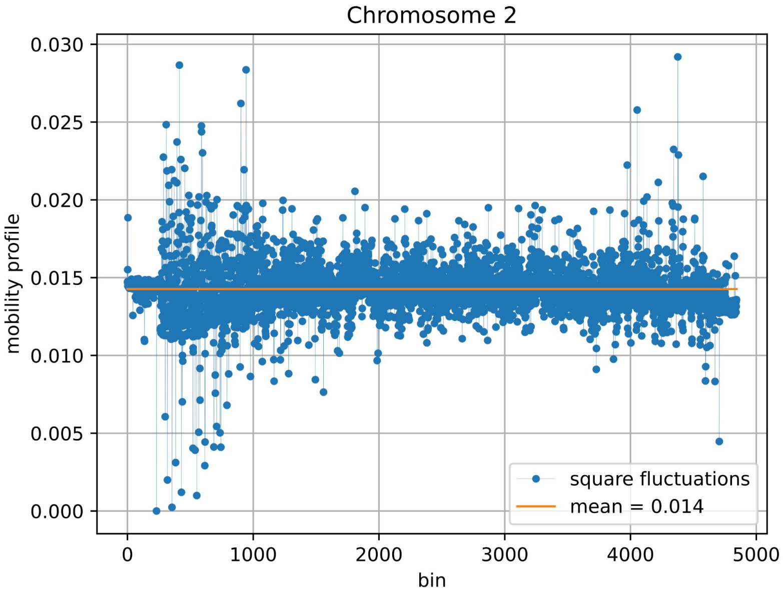
Covariance matrix

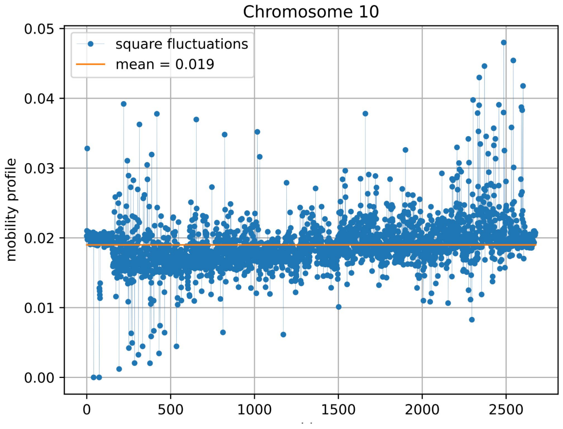
$$cov = \sum_k \frac{u_k\, u_k^\top}{\lambda_k}$$

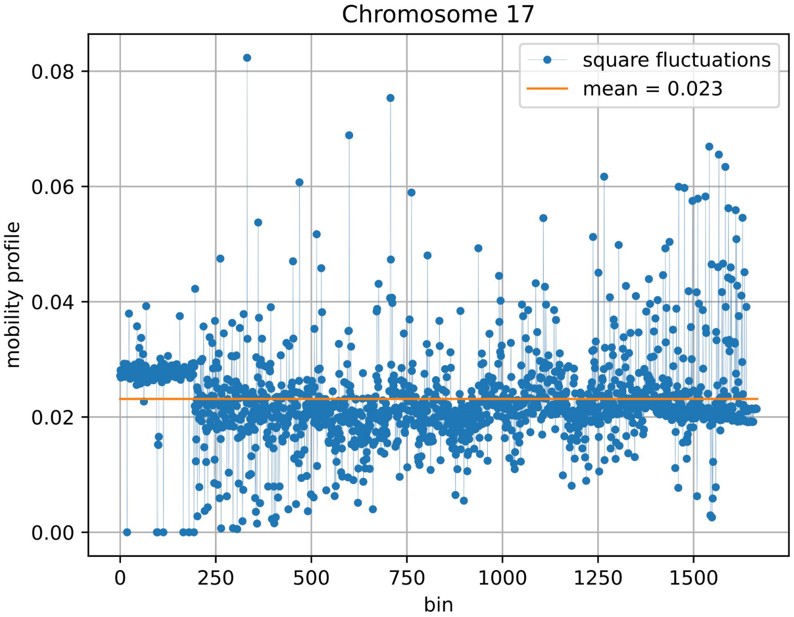
We only need diagonal, so

$$\sum_k \frac{u_k \circ u_k}{\lambda_k}$$

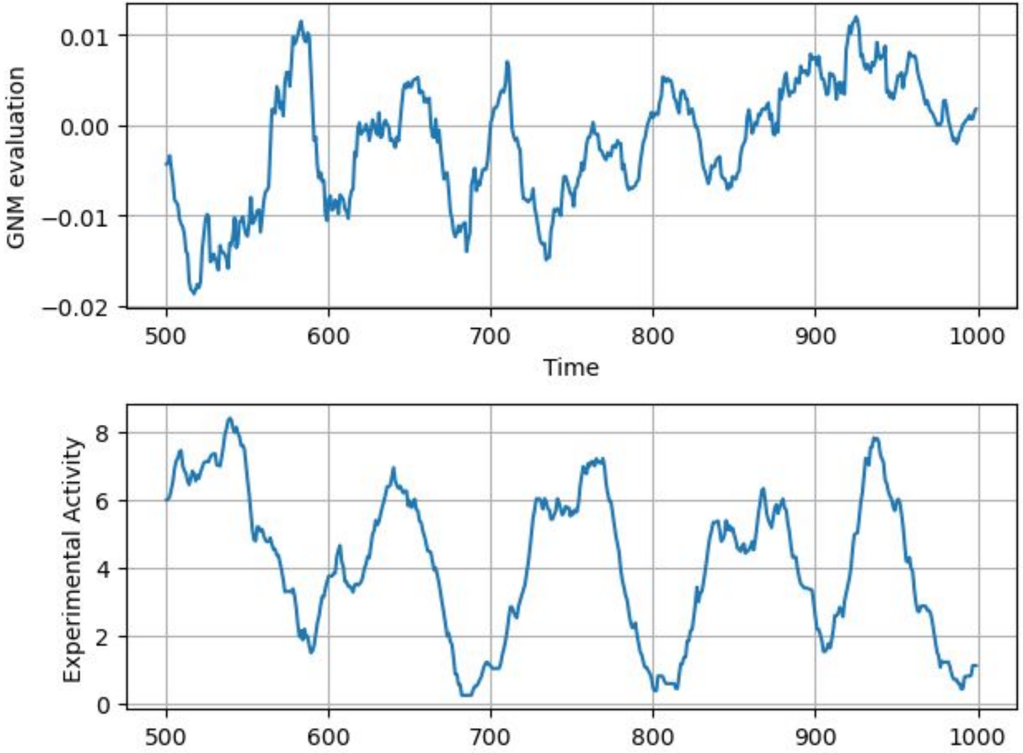
Predicted mean-square fluctuation (MSF)

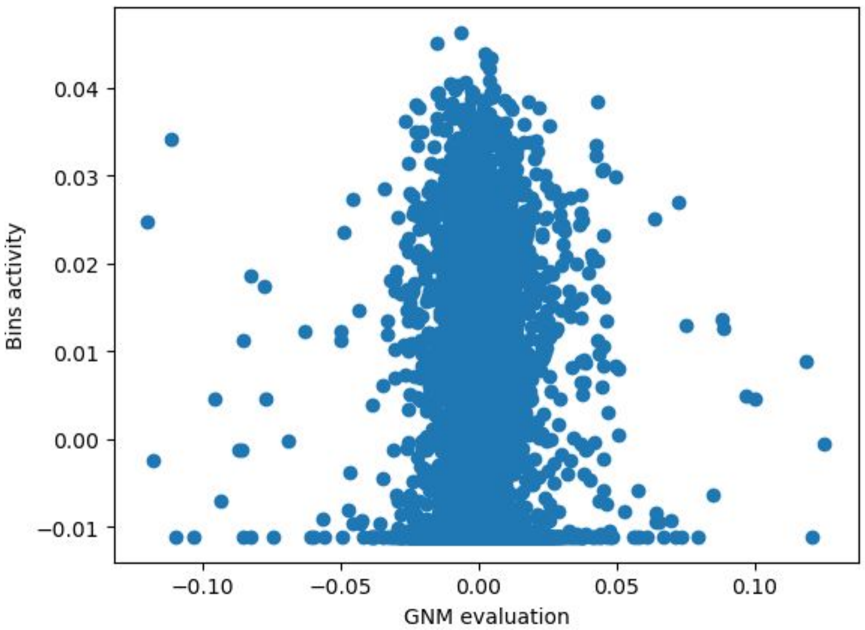

Chromosome 2
mobility profile
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
1000
2000
3000
4000
5000
bin
square fluctuations
mean = 0.014


Chromosome 10
mobility profile
0.05
0.04
0.03
0.02
0.01
0.00
0
500
1000
1500
2000
2500
square fluctuations
mean = 0.019

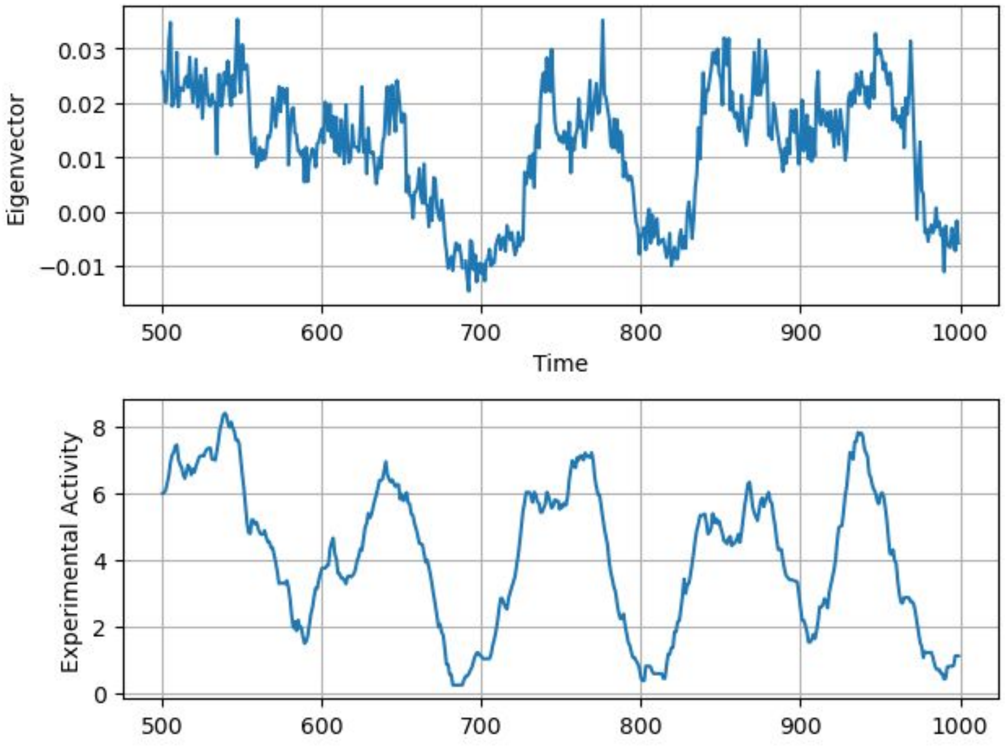

Chromosome 17
mobility profile
0.08
0.06
0.04
0.02
0.00
0
250
500
750
1000
1250
1500
bin
square fluctuations
mean = 0.023

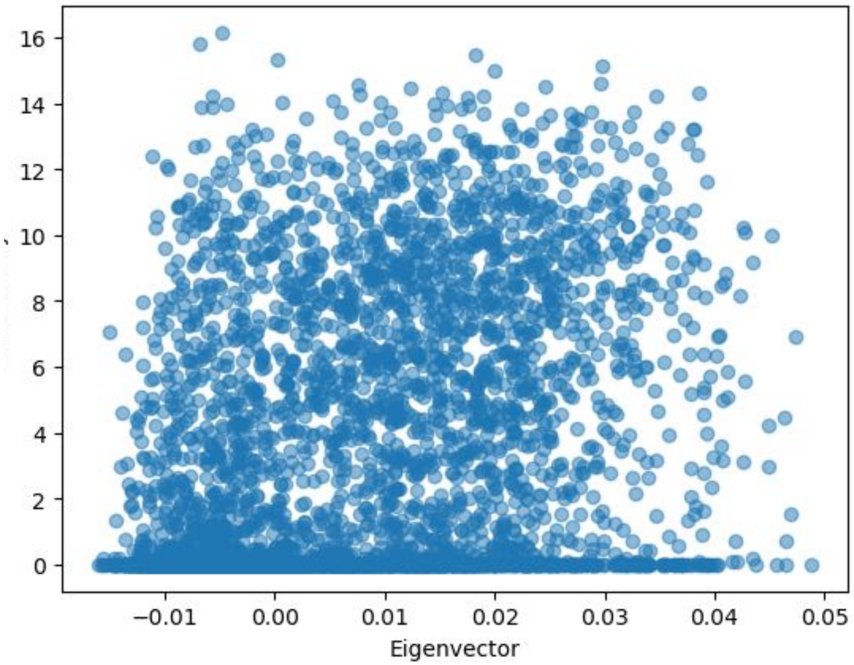
Gaussian Network Model Evaluation is hardly related with bin activity



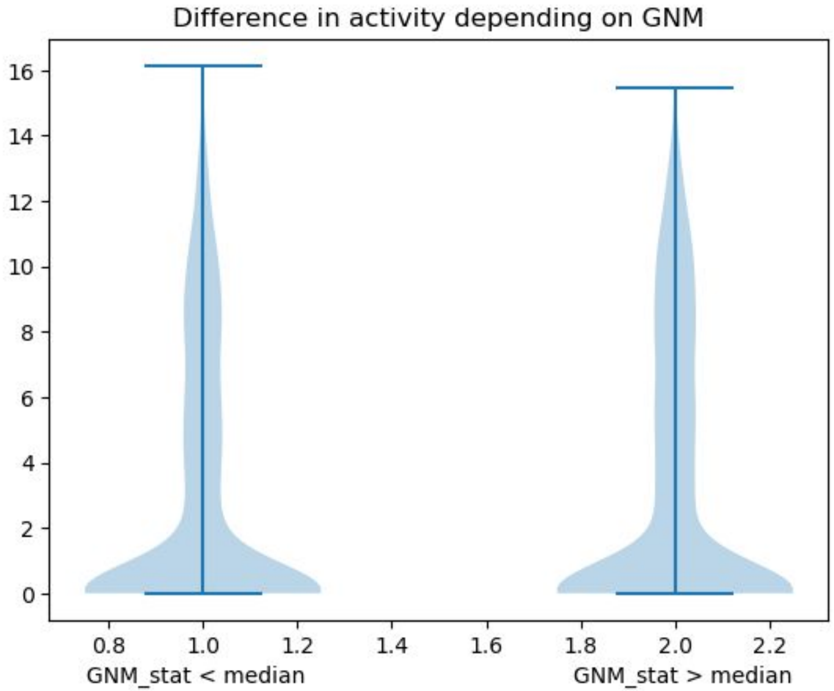


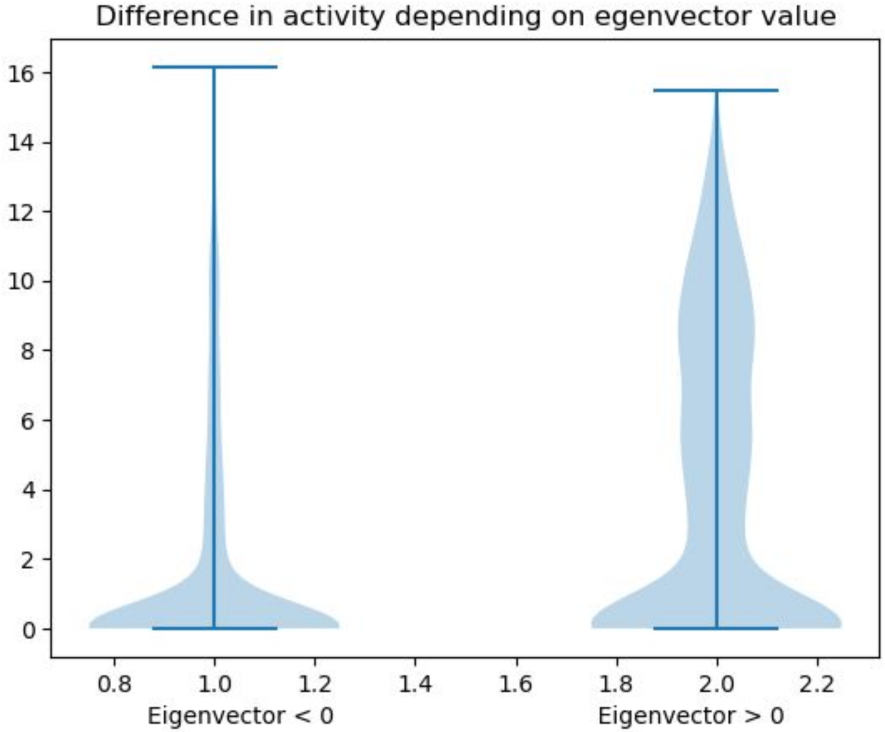
Eigenvector provides valuable information about gene activity





Eigenvector sign is determined by gene activity





$p < 10^{-26}$

Conclusions

The First Eigenvector is well-related with gene activity, that is complement with other Hi-C studies

Gaussian Network Model Evaluation shows no capacity to distinguish compartments

Bad performance could be caused by noise from other eigenvectors or some problems within the algorithm