PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Hi-C Chromatin Interaction Networks Predict Co-expression in the Mouse Cortex

Sepideh Babaei[1☯], Ahmed Mahfouz[1,2☯], Marc Hulsman[1,3], Boudewijn P. F. Lelieveldt[2,4], Jeroen de Ridder[1]*, Marcel Reinders[1]*

1 Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands, **2** Division of Image Processing, Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands, **3** Department of Clinical Genetics, VU University Medical Center, Amsterdam, The Netherlands, **4** Department of Intelligent Systems, Delft University of Technology, Delft, The Netherlands

☯ These authors contributed equally to this work.
* J.deRidder@tudelft.nl (JDR); M.J.T.Reinders@tudelft.nl (MR)

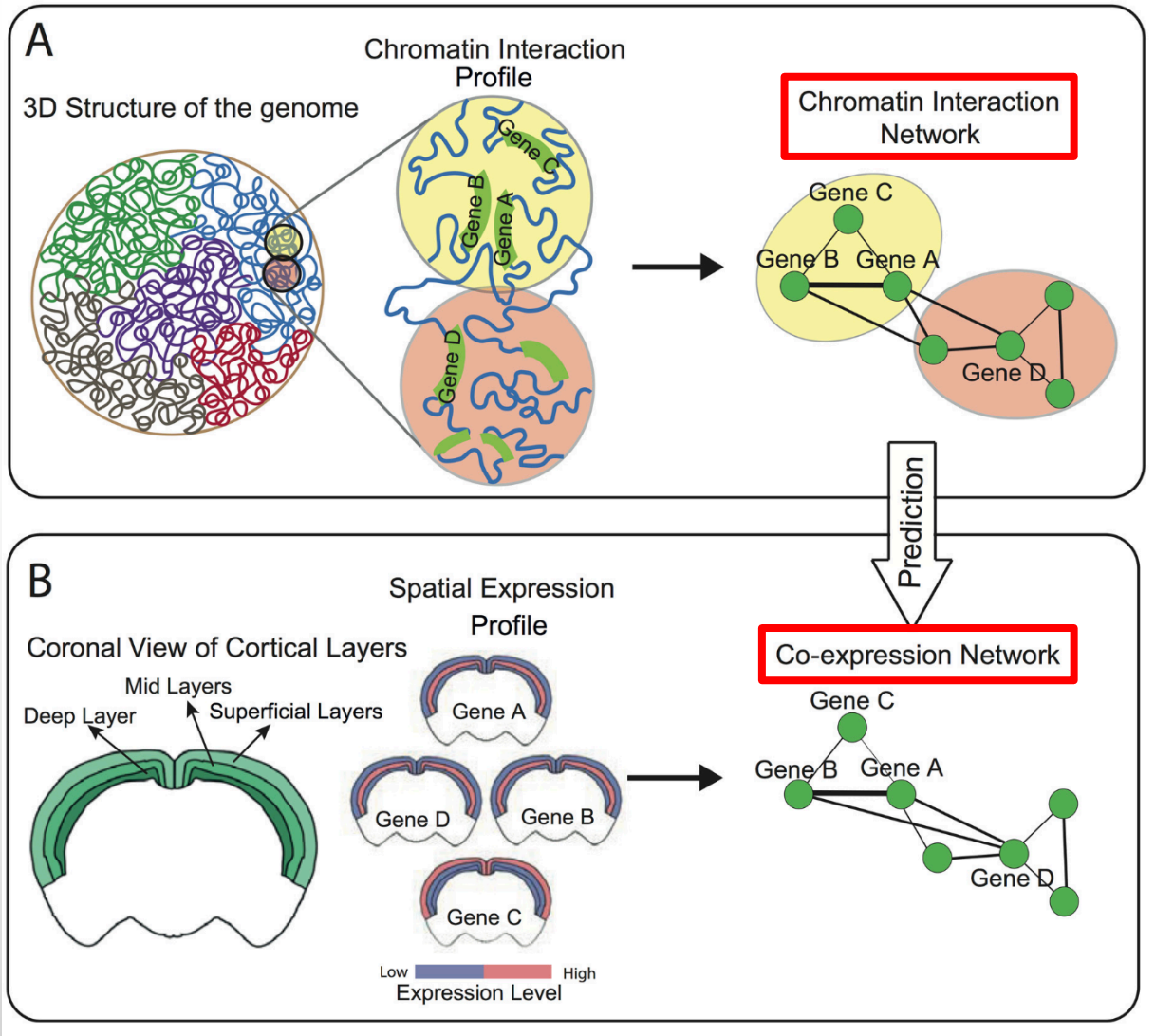
Shaurya Jauhari
Mora Lab, GMU.
(shauryajauhari@gzhmu.edu.cn)

OBJECTIVES

- CO-EXPRESSED ARE CO-FUNCTIONAL AND CO-LOCAL
- CO-EXPRESSION VIA CHROMATIN INTERACTIONS

PREFACE

- 3D GENOME ORGANIZATION
- MAPPING CO-EXPRESSED GENOMIC REGIONS (HI-C DATA)
- STRONG CORRELATION BETWEEN CHROMATIN INTERACTIONS AND GENE CO-EXPRESSION
  - **PREDICTION**
- **SCALE AWARE TOPOLOGICAL MEASURES : MULTI-ZOOM LEVELS**
  - **INTERACTIONS BETWEEN GENOMIC REGIONS (SMALL), GENES (MEDIUM), CHROMATIN COMPARTMENTS (LARGE)**


A
3D Structure of the genome
Chromatin Interaction Profile
Gene C
Gene B
Gene A
Gene D
Chromatin Interaction Network
Gene C
Gene B
Gene A
Gene D
Prediction
B
Coronal View of Cortical Layers
Mid Layers
Deep Layer
Superficial Layers
Spatial Expression Profile
Gene A
Gene D
Gene B
Gene C
Low
High
Expression Level
Co-expression Network
Gene C
Gene B
Gene A
Gene D

NETWORK TOPOLOGY

- GENOMIC LOCUS -> NODE -> NON-OVERLAPPING BINS OF A PARTICULAR RESOLUTION
- CHROMATIN INTERACTIONS -> LINKS/ EDGES -> HI-C SCORES

- SCALE AWARE TOPOLOGICAL MEASURES (HULSMAN ET AL.)
  - DIFFUSION KERNELS, NETWORK SMOOTHING PROCESS
  - DIFFUSION STRENGTH β *DETERMINES THE SCALE AT WHICH THE NETWORK IS CONSIDERED*

$$\beta = \frac{2^{6b} - 1}{2^6 - 1} * (10 - 0.0001) + 0.0001$$

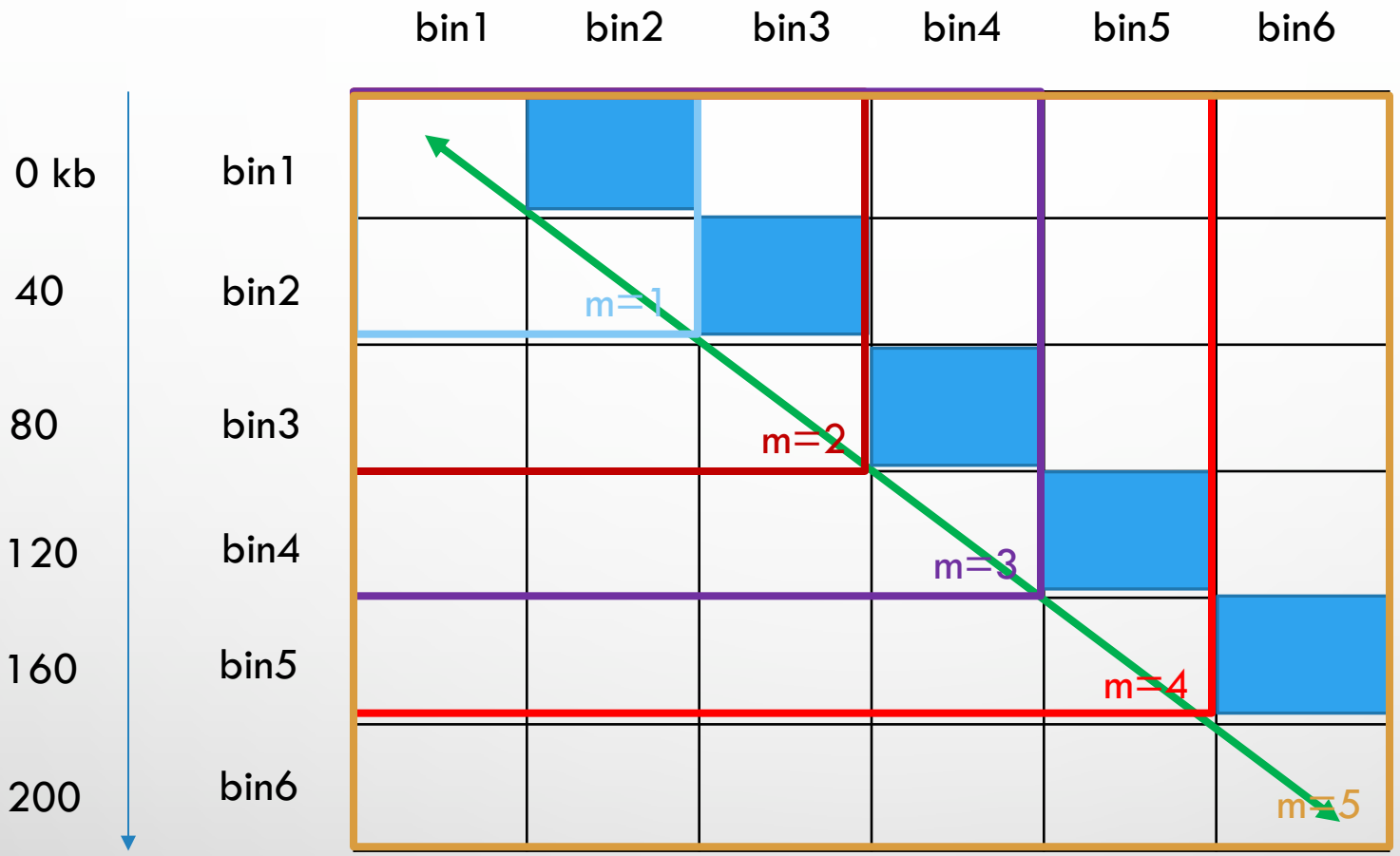
  - WITH $b = 0.0, \ldots, 1.0$ IN 10 STEPS RESULTING β:[0.0001, 0.09, 0.24, 0.47, 0.8, 1.4, 2.3, 3.8, 6.2, 10].

DATA

- THE ALLEN MOUSE BRAIN ATLAS (ABA)
  - 8–WEEK OLD ADULT C57BL/6J MALE MOUSE BRAIN
- INTRA-CHROMOSOMAL HI-C DATA (SHEN ET AL.)
  - BINS: 40KB GENOMIC SEGMENTS
  - CELL SPECIFIC HI-C MEASUREMENTS
  - RANK BASED NORMALIZATION OF THE HI-C SCORE CORRESPONDING TO A DISTANCE
    - CONFRONTING GENOMIC DISTANCE BIAS (TWO REGIONS CLOSE TO EACH OTHER IN THE LINEAR GENOME HAVE HIGHER CONTACT FREQUENCY DESPITE THEIR SPATIAL ORGANIZATION).

RANK BASED NORMALIZATION

- THE NORMALIZED HI-C SCORE $\hat{c}_{ij}$ IS DEFINED AS THE RANK OF c_{ij} IN THE VECTOR C^d , WHERE c_{ij} IS THE HI-C CONTACT BETWEEN BIN *I* AND *J* WITH GENOMIC DISTANCE OF D BASE PAIRS (BP).
- THE VECTOR C^d IS THE M^{th} SUPER-DIAGONAL OF THE HI-C CONTACT MATRIX WITH $\mathrm{M} = \frac{D}{binsize}$ WHICH CONTAINS HI-C SCORES BETWEEN ALL BIN PAIRS THAT HAVE SAME GENOMIC DISTANCE *D*.


bin1
bin2
bin3
bin4
bin5
bin6
0 kb
40
80
120
160
200
bin1
bin2
bin3
bin4
bin5
bin6
m=1
m=2
m=3
m=4
m=5

DATA (CONTD.)

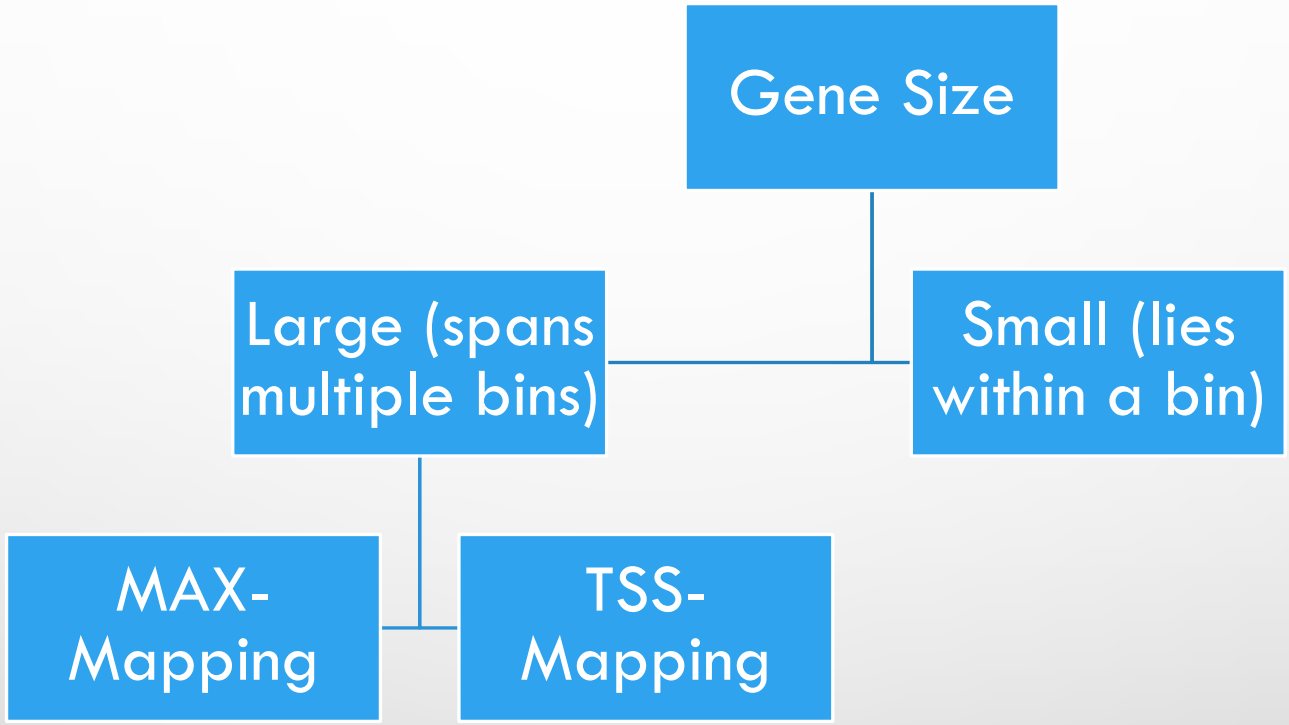
- MULTI-RESOLUTION HI-C DATA
  - OBJECTIVE : PREDICTION OF GENE CO-EXPRESSION PATTERNS
  - BIN-BASED HI-C MATRIX -> GENE-BASED HI-C MATRIX
    - ALSO AT MULTI-ZOOM LEVELS

DATA (CONTD.)

- CHROMATIN INTERACTION NETWORK
- HI-C INTERACTIONS BETWEEN EACH GENE-PAIR ~ HI-C INTERACTIONS BETWEEN CORRESPONDING BINS
    - SOME GENES MAY SPAN MULTIPLE BINS DEPENDING ON THE GENE SIZE

Variable Hi-C Matrices

DETERMINING HI-C INTERACTION



DETERMINING HI-C INTERACTION (LARGER GENES THAN BINS CAN ACCOMODATE)

- MAX-MAPPING
  - A LINK (EDGE) IS DEFINED AS THE MAXIMUM HI-C VALUE AMONG ALL POSSIBLE INTERACTIONS
  - $\hat{h}_{xy} = max_{i \subset x, j \subset y}(\hat{h}_{ij})$, WHERE (X,Y) SYMBOLIZES A GENE PAIR.

- TSS-MAPPING
  - A LINK (EDGE) IS DEFINED AS THE BIN-PAIR THAT CONTAINS TSS OF TWO GENES
  - $\hat{h}_{xy} = \hat{h}_{ij}$, WHERE TSS(X) $\subset$ I AND TSS(Y) $\subset$ J.

CHROMATIN INTERACTION NETWORK (CIN)

- WEIGHTED GENE-BASED HI-C MATRIX TO UNWEIGHTED MATRIX
  - RETAINING ONLY TRANSACTIONS THAT EXCEED THE 90TH PERCENTILE OF ALL HI-C SCORES ACROSS CHROMOSOMES AT THE CORRESPONDING BIN SIZE
- ONE CIN PER CHROMOSOME PER RESOLUTION
  - $H_{chr}^{R} = (G, I_H)$, WHERE 'G' REPRESENTS THE SET OF NODES CORRESPONDING TO BINS AND 'I_H' REPRESENTS THE SET OF LINKS CORRESPONDING TO HI-C INTERACTIONS BETWEEN GENES THAT EXCEED THE 90TH PERCENTILE OF ALL HI-C SCORES ACROSS CHROMOSOMES AT A RESOLUTION R.

CIN TOLOPOLOGY

- THE NODE-BASED TOPOLOGY MEASURES
  - SHORTEST PATH LENGTH, JACCARD INDEX, DEGREE (AND CLOSENESS) CENTRALITY, BETWEENNESS CENTRALITY, AND CLUSTERING COEFFICIENT
  - CONVERTED TO LINK-BASED TOPOLOGY MEASURES
    - BECAUSE FOR CLASSIFICATION/ PREDICTION (ML APPLICATION), THE MEASURES MUST BE REFLECTED AS FEATURES.
- CONVERSION
  - TAKE AVERAGE AND DIFFERENCE OF THE GENE-BASED MEASURE FOR EACH GENE PAIR.
  - EG. $\{\,|BC(x) - BC(y)|, \frac{1}{2}(BC(x) - BC(y)\}$

Neo4j

Table 1. Topological measures.

Measure	Description	Scale-aware version
Shortest Path	The minimum number of vertices connecting node x and y, $s(x, y)$	$s^{\beta}(x,y) = -\log(K^{\beta}_{x,y})$
Jaccard Index	The proportion of shared nodes between x and y relative to the total number of nodes connected to x or y, $J(x,y) = \frac{n(x) \cap n(y)}{n(x) \cup n(y)}$	$J^{\beta}(x,y) = \frac{\sum_i min(K^{\beta}_{x,i}, K^{\beta}_{i,y})}{\sum_i max(K^{\beta}_{x,i}, K^{\beta}_{i,y})}$
Degree & closeness Centrality	The degree centrality reflects the connectivity of a node in terms of the number of edges connected to it, *deg* (x) and closeness centrality reflects the farness of a node x, by summing the shortest path distances to all other nodes, $c(x) = \frac{1}{\sum_{i \setminus x} s(x,i)}$	$c^{\beta}(x) = 1 - K^{\beta}_{x,x}$
Betweenness Centrality	The number of shortest paths that pass through a node, $b(x) = \sum_{i,j \setminus x} \frac{q_{ij}(x)}{q_{ij}}$ where q_{ij} is the number of shortest paths between nodes i and j, and $q_{ij}(x)$ the number of those paths that pass through x	$b^{\beta}(z) = \frac{1}{N^2} \sum_{x,y} (s^{\beta}(x,y) - (s^{\beta}(x,z) + s^{\beta}(z,y)))$
Clustering Coefficient	The number of edges between its direct neighbors including itself, divided by the maximum number of possible edges, $cc(x) = \frac{2\lvert e_x \rvert}{deg(x)(deg(x)-1)}$	$cc^{\beta}(x) = \sum_{i \setminus x} K^{\beta}_{x,i} J^{\beta}(x,i)$

N is the set of all nodes in the network, and n is the number of nodes. (x, y) is a link between nodes x and y, $(x, y \in N)$. $a(x, y)$ is the connection status between x and y: $a(x, y) = 1$ when link (x, y) exists; $a(x, y) = 0$ otherwise. Scale-aware versions are based on diffusion kernel where $K^{\beta} = e^{\beta(A - D)}$, A is the adjacency matrix and D is the degree matrix of the network. The diffusion level β determines the scale. $K^{\beta}(x, y)$ is the diffusion strength between node x and y.

doi:10.1371/journal.pcbi.1004221.t001

LINK METADATA (8 FEATURES)

1. Clustering Coefficient (Difference, Average)

2. Betweenness Centrality (Difference, Average)

Gene X → Gene Y

3. Degree Centrality (Difference, Average)

4. Closeness Centrality (Difference, Average)

HIGHLY CO-EXPRESSED GENES ARE SPATIALLY CO-LOCALIZED

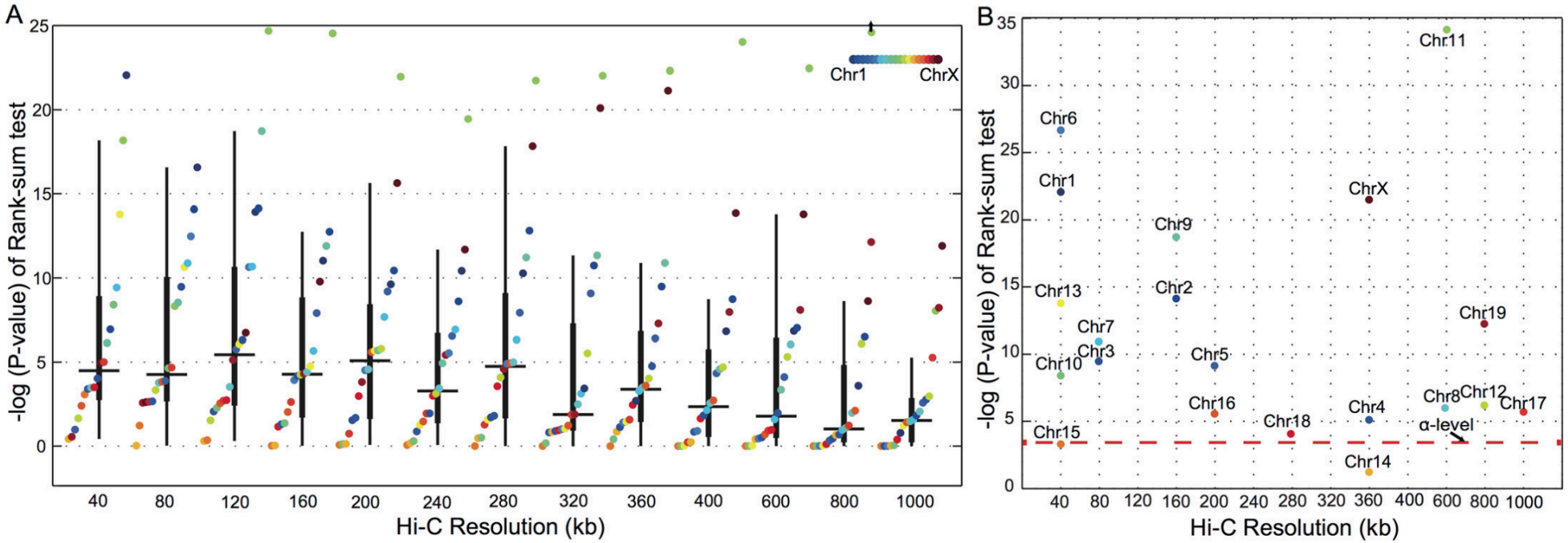


Fig 2. Co-expressed genes are co-localized in 3D structure of the genome. (A) Assessment of the enrichment of Hi-C interactions between strongly co-expressed gene-pairs compared to gene-pairs with no co-expression across different Hi-C resolutions. The y-axis indicates $-log_{10}(p - value)$ of the one-tailed Wilcoxon rank-sum test used for the enrichment analysis. Hi-C interactions were mapped to genes using the MAX-mapping method. (B) Overview of the Hi-C resolution at which Hi-C interactions are most significantly associated with co-expressed gene-pairs for each chromosome. In each box, the horizontal line represents the median. The thick vertical line represents the interval of $q_1 = 25^{th}$ and $q_3 = 75^{th}$ percentiles. The thin vertical line represents the interval of $q_3 + 1.5(q_3 - q_1)$ and $q_1 - 1.5(q_3 - q_1)$.

doi:10.1371/journal.pcbi.1004221.g002

PREDICTING CO-EXPRESSIONS VIA CHROMATIN INTERACTIONS

- REQUISITES
  - HI-C CONTACT MATRIX
  - CO-EXPRESSION MATRIX

- INDICATORS
  - CHROMATIN INTERACTION ADJACENCY DOESN'T NECESSARILY IMPLY CO-EXPRESSION
  - CHARACTERIZATION OF LONG RANGE INTERACTIONS IS IMPORTANT

CHROMATIN INTERACTION AND CO-EXPRESSION HARBOR INDIRECT RELATIONSHIP

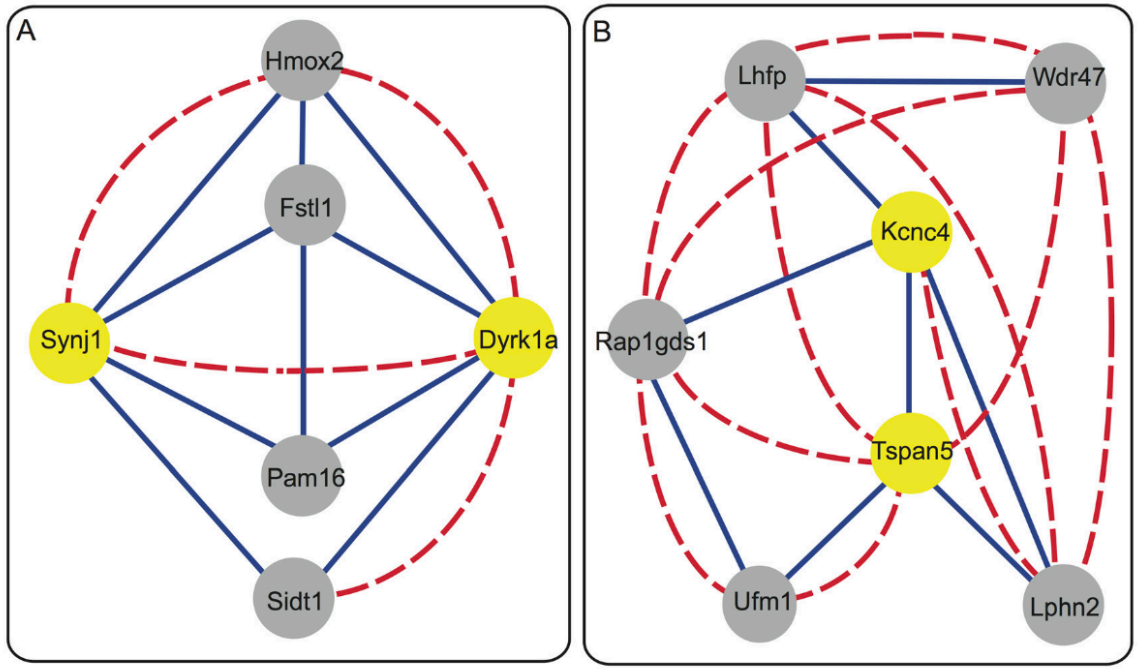


Fig 3. Chromatin interactions of gene-pairs in the CIN at 200*kb* resolution. (A) *Synj*1-*Dykr*1*a* (yellow nodes) in Chromosome 16 are co-expressed (dashed red link) but their corresponding genomic loci do not interact frequently (no blue link). Both genes have strong chromatin interactions with 4 other genes (grey nodes) resulting a high Jaccard index between them. (B) *Kcnc*4-*Tapan*5 (yellow nodes) in Chromosome 3 directly interact (solid blue line) but they are not strongly co-expressed (no dashed red line). This direct chromatin interaction explains the strong co-expression between other gene-pairs in their neighbourhood, such as *Wdr*47-*Tapan*5 and *Wdr*47-*Rap*1*gds*1, which are not directly connected in the CIN themselves (no solid blue line). The betweenness centrality measure of the link between *Kcnc*4-*Tapan*5 can describe the strong co-expression between their neighbouring genes. Chromatin interaction and co-expression are shown by solid blue and dashed red links, respectively.

doi:10.1371/journal.pcbi.1004221.g003

STMS ENHANCE PREDICTION PERFORMANCE

- FEATURES DESCRIBED BY LINK ATTRIBUTES
  - DEGREE CENTRALITY
  - CLOSENESS CENTRALITY
  - BETWEENNESS CENTRALITY
  - CLUSTERING COEFFICIENT

- 8 SCALE AWARE TOPOLOGICAL MEASURES:
  - ACROSS 10 SCALES (β VALUES)
  - ACROSS 10 RESOLUTIONS CHROMATIN INTERACTION NETWORKS

800 features

STMS ENHANCE PREDICTION PERFORMANCE (CONTD.)

- CIN TOPOLOGY WAS DESCRIBED AT MULTIPLE TOPOLOGICAL SCALES USING STMS
- STMS OF CIN OF EACH HI-C RESOLUTION WAS CALCULATED SEPARATELY AND THEN CONCATENATED TOGETHER
- STMS BETTER PREDICT GENE-PAIRS AS CO-EXPRESSED (POSITIVE CLASS : CORRELATION ABOVE 50 PERCENTILE **OF ALL CORRELATIONS ACROSS ALL CHROMOSOMES**) AND NOT CO-EXPRESSED (NEGATIVE CLASS: CORRELATION BELOW 50 PERCENTILE)

TAKEAWAYS

- CHROMATIN INTERACTIONS ARE CHROMOSOME SPECIFIC
- CONSIDERING NETWORK AT DIFFERENT SCALES ALSO HIGHLIGHTS THE HIERARCHY OF INTERACTIONS IN CHROMATIN FOLDING (SANDHU ET AL.)