Topologically associating domains of chromatin: methods and tools for calling

Part 1

Svyatoslav Sidorov[1]

[1]The Dobzhansky Center for Genome Bioinformatics
St. Petersburg State University

Group meeting at BI

Outline

Outline

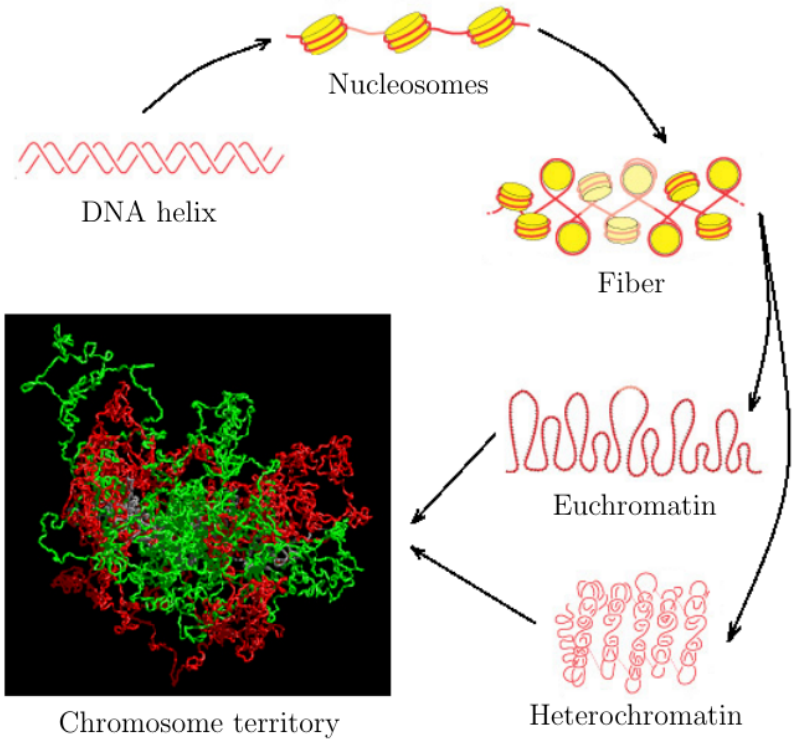
Outline

Outline

Outline

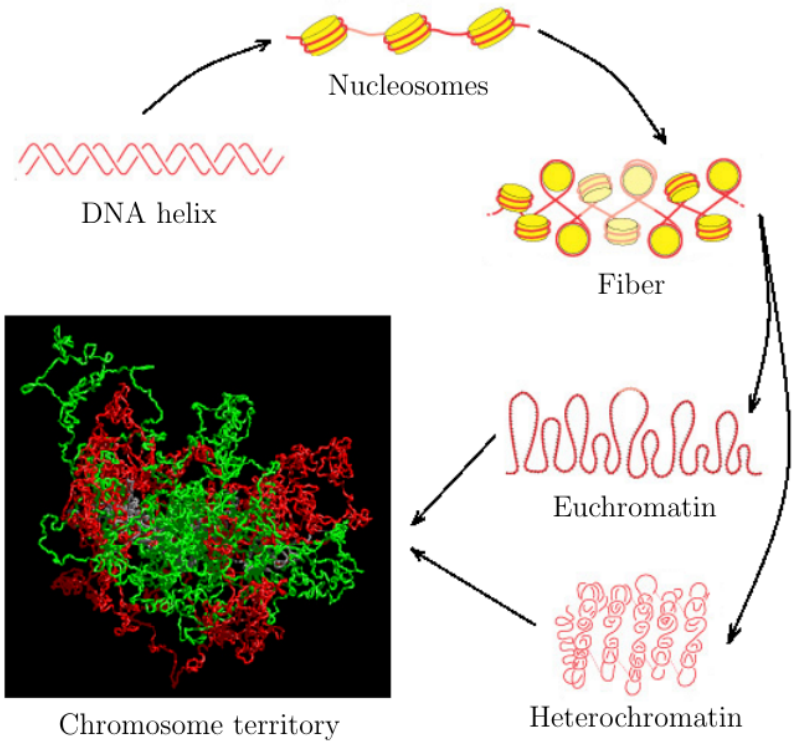
Outline

Introduction



Alberts B. et al. 2004. Essential Cell Biology, 2 ed.; Koch T. A. et al.

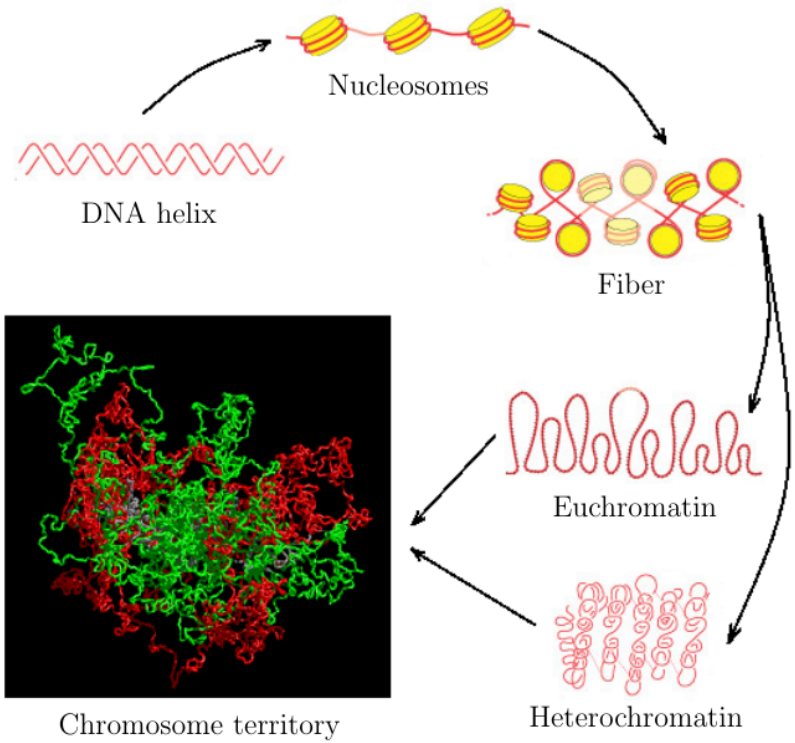
Introduction



Alberts B. et al. 2004. Essential Cell Biology, 2 ed.; Koch T. A. et al.

- **Question:** How is chromatin folded within euchromatin and heterochromatin compartments?

Introduction

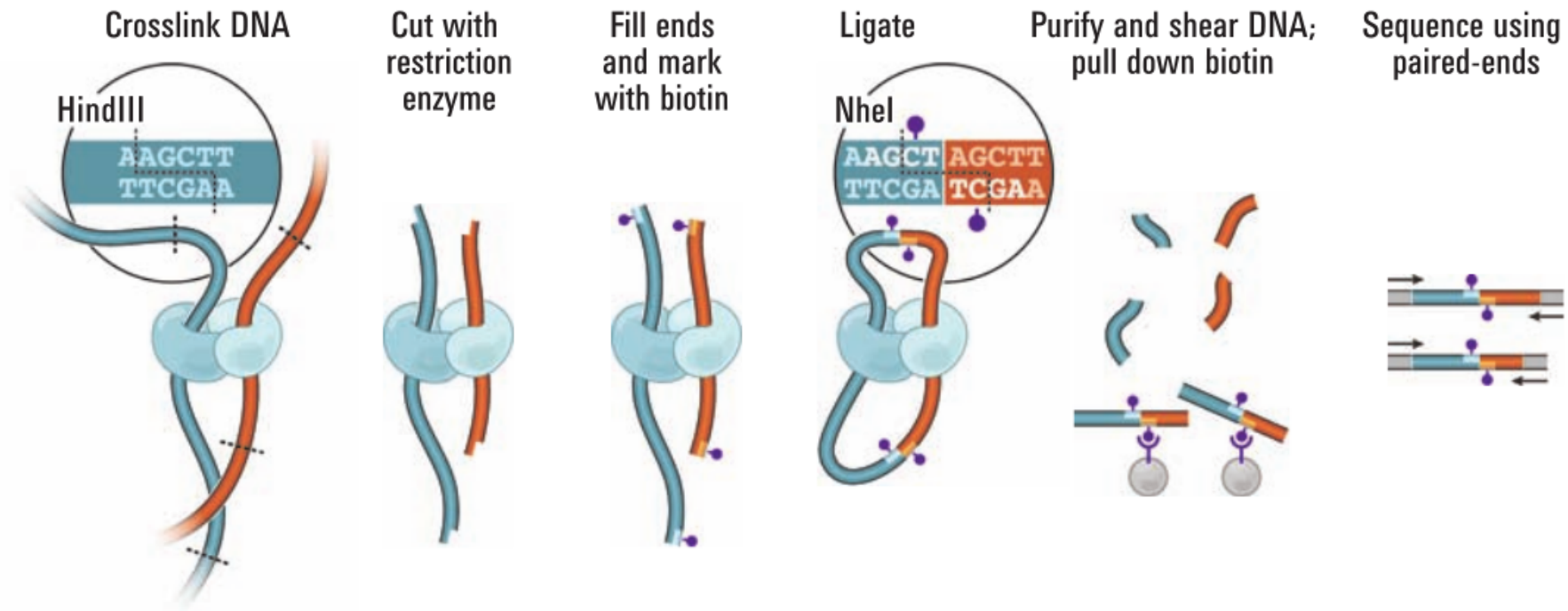


Alberts B. et al. 2004. Essential Cell Biology, 2 ed.; Koch T. A. et al.

- **Question:** How is chromatin folded within euchromatin and heterochromatin compartments?
- **The answer** came with the development of chromatin conformation capture methods (3C, 2002; 4C, 2006; 5C, 2006; **Hi-C**, 2009).

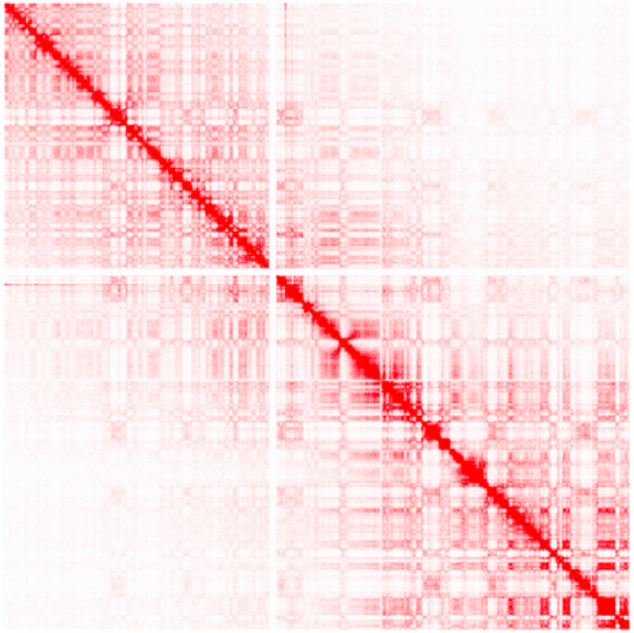
Introduction

Hi-C experiment scheme:



Lieberman-Aiden et al., 2009

Introduction

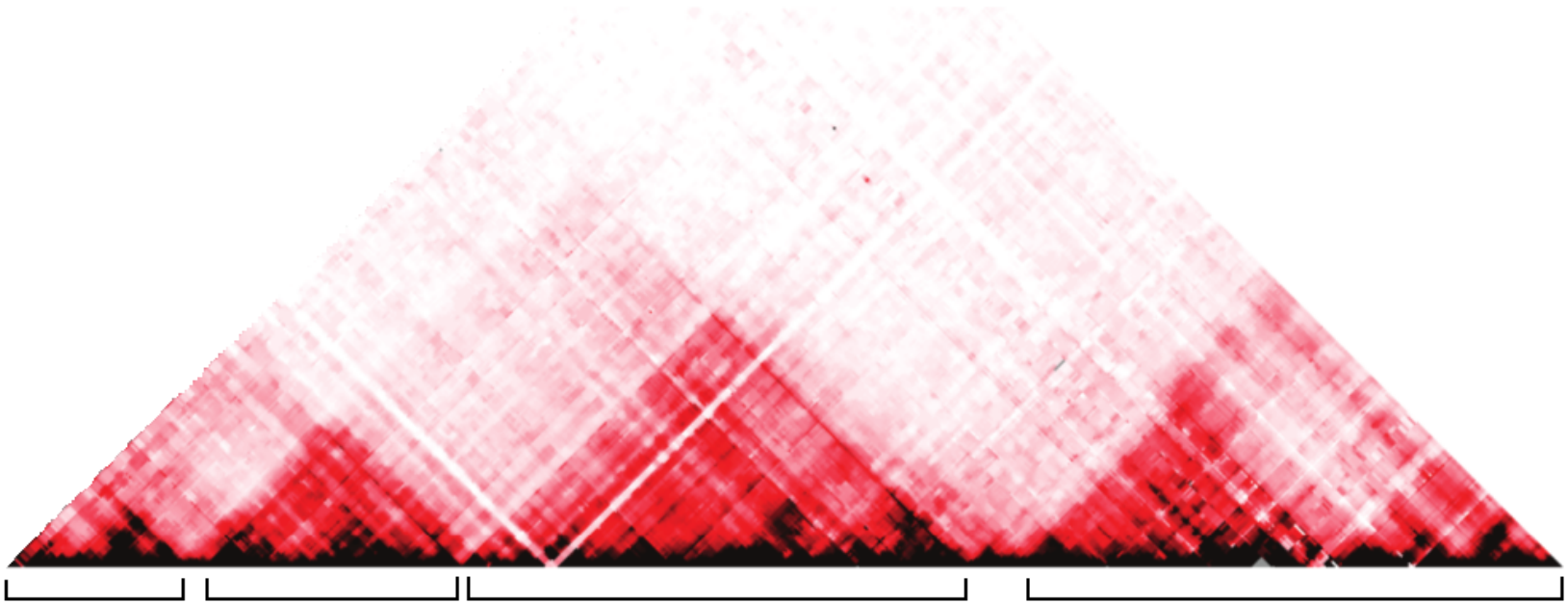


HOMER tool website

- Chromosome is split into r bp bins (r is called **contact matrix resolution**).
- **Contact matrix** C is built: $C(i, j) \equiv C(j, i)$ is a number of paired-end reads such that one read was mapped into bin i and the other read was mapped into bin j. Contact matrix is usually represented as a heatmap.

Outline

Topologically associating domains



Self-interacting domains can be seen on the main diagonal of a contact matrix (Dekker et al., 2013, adapted).

Topologically associating domains

- Dixon et al., 2012 found self-interacting domains in human and mouse using Hi-C data.

LETTER

doi:10.1038/nature11082

Topological domains in mammalian genomes identified by analysis of chromatin interactions

Jesse R. Dixon[1,2,3], Siddarth Selvaraj[1,4], Feng Yue[1], Audrey Kim[1], Yan Li[1], Yin Shen[1], Ming Hu[5], Jun S. Liu[5] & Bing Ren[1,6]

Topologically associating domains

- Dixon et al., 2012 found self-interacting domains in human and mouse using Hi-C data.

LETTER

doi:10.1038/nature11082

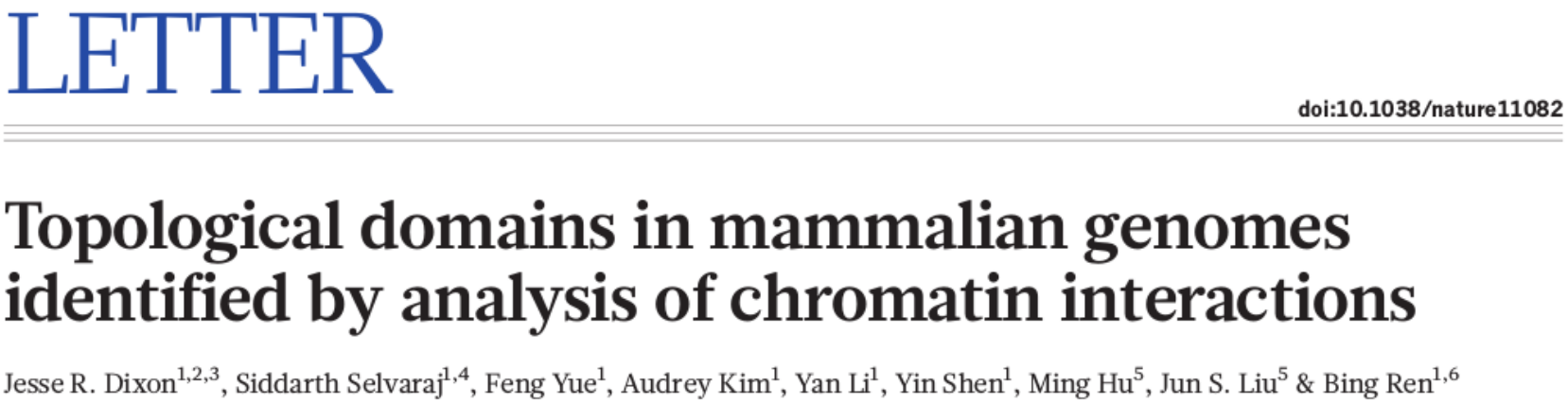
Topological domains in mammalian genomes identified by analysis of chromatin interactions

Jesse R. Dixon[1,2,3], Siddarth Selvaraj[1,4], Feng Yue[1], Audrey Kim[1], Yan Li[1], Yin Shen[1], Ming Hu[5], Jun S. Liu[5] & Bing Ren[1,6]

- They called such domains **topologically associating domains (TADs)**. TAD is such a region that frequency of intra-TAD interactions is higher than inter-TAD interactions.

Topologically associating domains

- Dixon et al., 2012 found self-interacting domains in human and mouse using Hi-C data.



LETTER

doi:10.1038/nature11082

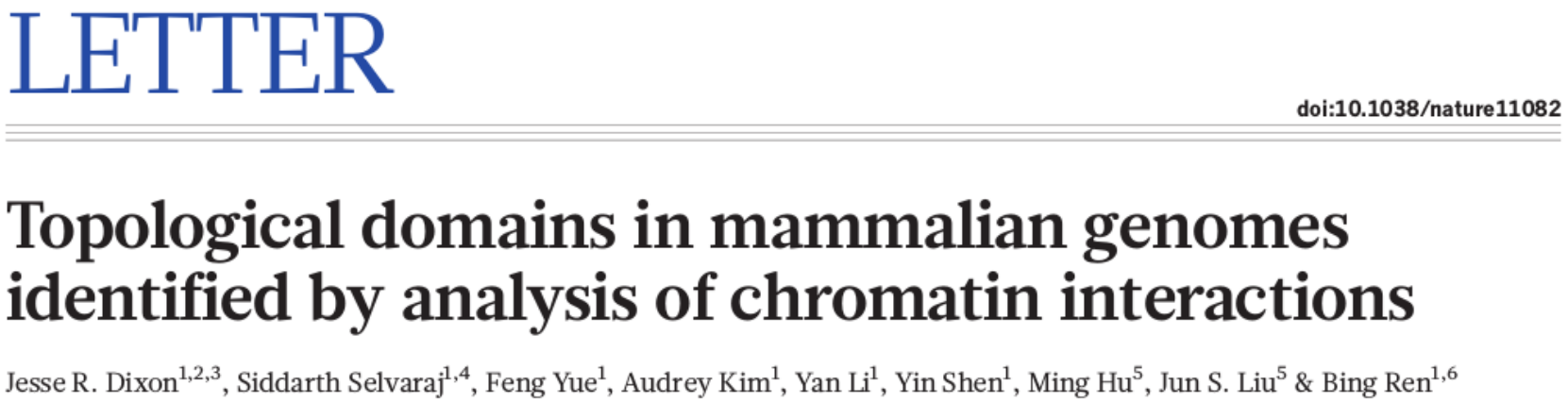
Topological domains in mammalian genomes identified by analysis of chromatin interactions

Jesse R. Dixon[1,2,3], Siddarth Selvaraj[1,4], Feng Yue[1], Audrey Kim[1], Yan Li[1], Yin Shen[1], Ming Hu[5], Jun S. Liu[5] & Bing Ren[1,6]

- They called such domains **topologically associating domains (TADs)**. TAD is such a region that frequency of intra-TAD interactions is higher than inter-TAD interactions.
- Similar domains were found in *Drosophila* genome in the same year: Sexton et al., 2012; Hou et al., 2012.

Topologically associating domains

- Dixon et al., 2012 found self-interacting domains in human and mouse using Hi-C data.



LETTER

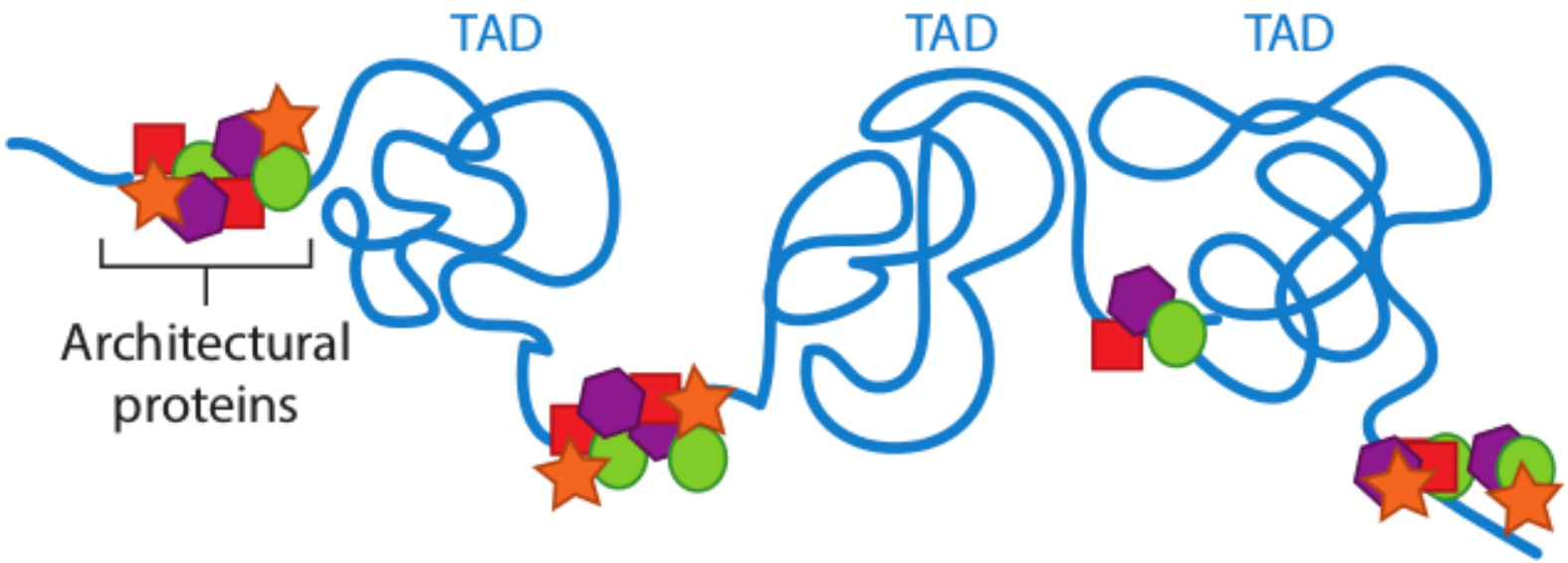
doi:10.1038/nature11082

Topological domains in mammalian genomes identified by analysis of chromatin interactions

Jesse R. Dixon[1,2,3], Siddarth Selvaraj[1,4], Feng Yue[1], Audrey Kim[1], Yan Li[1], Yin Shen[1], Ming Hu[5], Jun S. Liu[5] & Bing Ren[1,6]

- They called such domains **topologically associating domains (TADs)**. TAD is such a region that frequency of intra-TAD interactions is higher than inter-TAD interactions.
- Similar domains were found in *Drosophila* genome in the same year: Sexton et al., 2012; Hou et al., 2012.
- TADs were also found in the same year in mouse X chromosome by Nora et al., 2012.

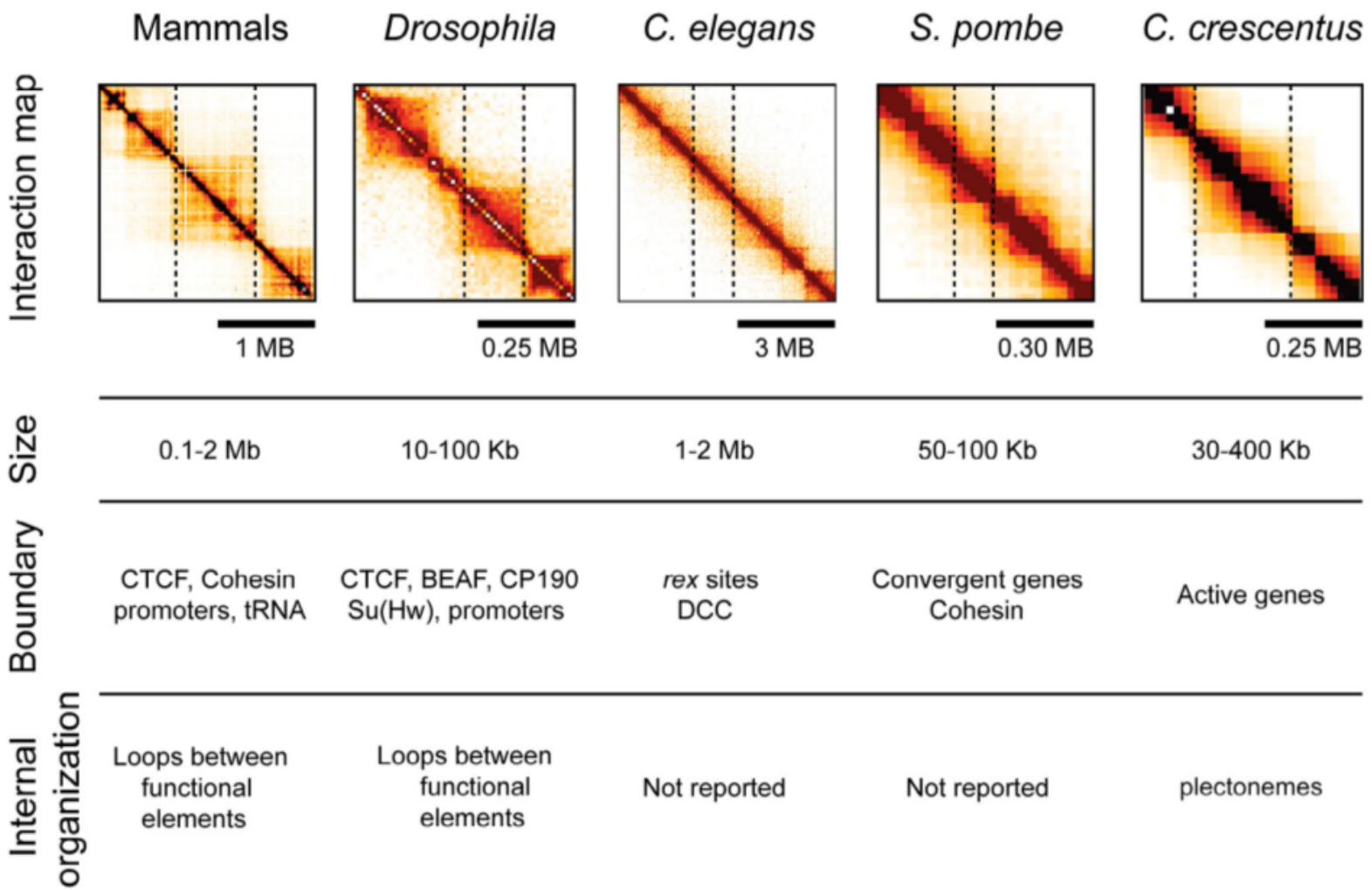
Topologically associating domains



Nguyen H. G. and Bosco G., 2015

- TADs are collections of many chromatin loops.
- TADs are separated by **TAD borders** (intervening chromatin).
- Mammalian TAD borders are enriched in active transcription, housekeeping genes, tRNA genes and SINE repeats, as well as binding sites for the architectural proteins CTCF and cohesin (Dekker J. and Heard E., 2015).

Topologically associating domains



TAD-like domains were found in several organisms in 2012 – 2015 (Dekker J. and Heard E., 2015, adapted).

Topologically associating domains

TADs as functional domains in mammals (Dekker J. and Heard E., 2015):

Topologically associating domains

TADs as functional domains in mammals (Dekker J. and Heard E., 2015):

- TADs are units of coordinated gene expression.

Topologically associating domains

TADs as functional domains in mammals (Dekker J. and Heard E., 2015):

- TADs are units of coordinated gene expression.
- Series of adjacent TADs correspond to replication domains.

Topologically associating domains

TADs as functional domains in mammals (Dekker J. and Heard E., 2015):

- TADs are units of coordinated gene expression.
- Series of adjacent TADs correspond to replication domains.
- Some TADs correspond to lamina-associated domains and other types of repressed chromatin.

Topologically associating domains

TADs as functional domains in mammals (Dekker J. and Heard E., 2015):

- TADs are units of coordinated gene expression.
- Series of adjacent TADs correspond to replication domains.
- Some TADs correspond to lamina-associated domains and other types of repressed chromatin.
- Mammalian TAD borders are to a significant extent conserved between different cell types, and even between mouse and human.

Topologically associating domains

TADs as functional domains in mammals (Dekker J. and Heard E., 2015):

- TADs are units of coordinated gene expression.
- Series of adjacent TADs correspond to replication domains.
- Some TADs correspond to lamina-associated domains and other types of repressed chromatin.
- Mammalian TAD borders are to a significant extent conserved between different cell types, and even between mouse and human.
- Cell type-specific enhancers make loops with promoters of corresponding genes predominantly within TADs.

Topologically associating domains

TADs as functional domains in mammals (Dekker J. and Heard E., 2015):

- TADs are units of coordinated gene expression.
- Series of adjacent TADs correspond to replication domains.
- Some TADs correspond to lamina-associated domains and other types of repressed chromatin.
- Mammalian TAD borders are to a significant extent conserved between different cell types, and even between mouse and human.
- Cell type-specific enhancers make loops with promoters of corresponding genes predominantly within TADs.
- Internal interaction patterns of TADs are highly cell type-specific.

Topologically associating domains

TADs as functional domains in mammals (Dekker J. and Heard E., 2015):

- TADs are units of coordinated gene expression.
- Series of adjacent TADs correspond to replication domains.
- Some TADs correspond to lamina-associated domains and other types of repressed chromatin.
- Mammalian TAD borders are to a significant extent conserved between different cell types, and even between mouse and human.
- Cell type-specific enhancers make loops with promoters of corresponding genes predominantly within TADs.
- Internal interaction patterns of TADs are highly cell type-specific.
- TADs have hierarchical folding and consist of **sub-TADs** (Cubeñas-Potts C. and Corces V. G., 2015; Rao et al., 2014).

Topologically associating domains

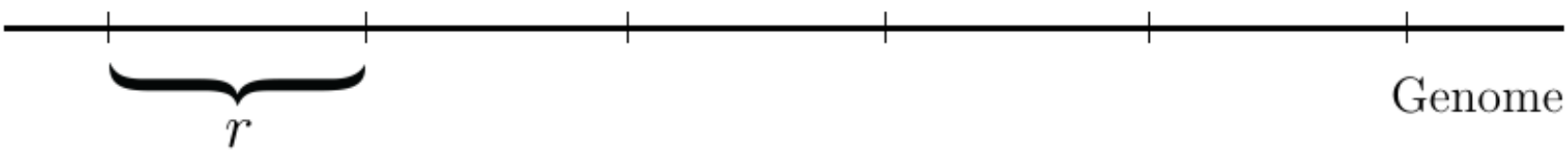
TADs as functional domains in mammals (Dekker J. and Heard E., 2015):

- TADs are units of coordinated gene expression.
- Series of adjacent TADs correspond to replication domains.
- Some TADs correspond to lamina-associated domains and other types of repressed chromatin.
- Mammalian TAD borders are to a significant extent conserved between different cell types, and even between mouse and human.
- Cell type-specific enhancers make loops with promoters of correspondent genes predominantly within TADs.
- Internal interaction patterns of TADs are highly cell type-specific.
- TADs have hierarchical folding and consist of **sub-TADs** (Cubeñas-Potts C. and Corces V. G., 2015; Rao et al., 2014).

Self-interacting domains in other organisms can have different functions (Dekker J. and Heard E., 2015).

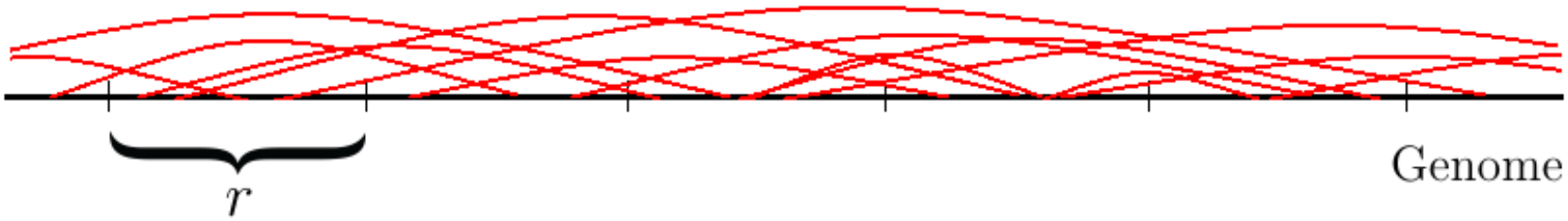
Outline

Directionality index



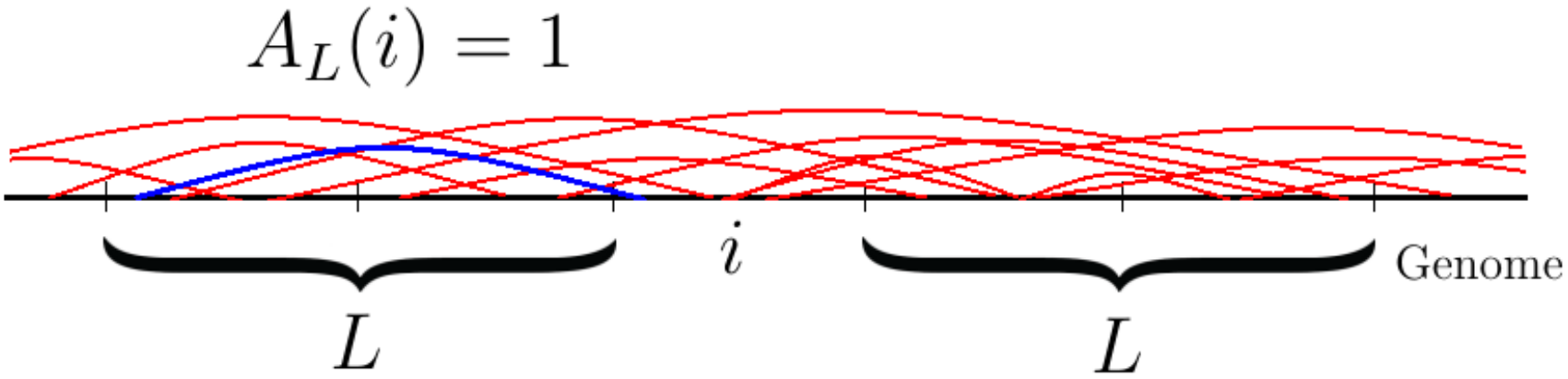
- Let's partition each chromosome into *r* bp bins, where *r* is a contact matrix resolution.

Directionality index



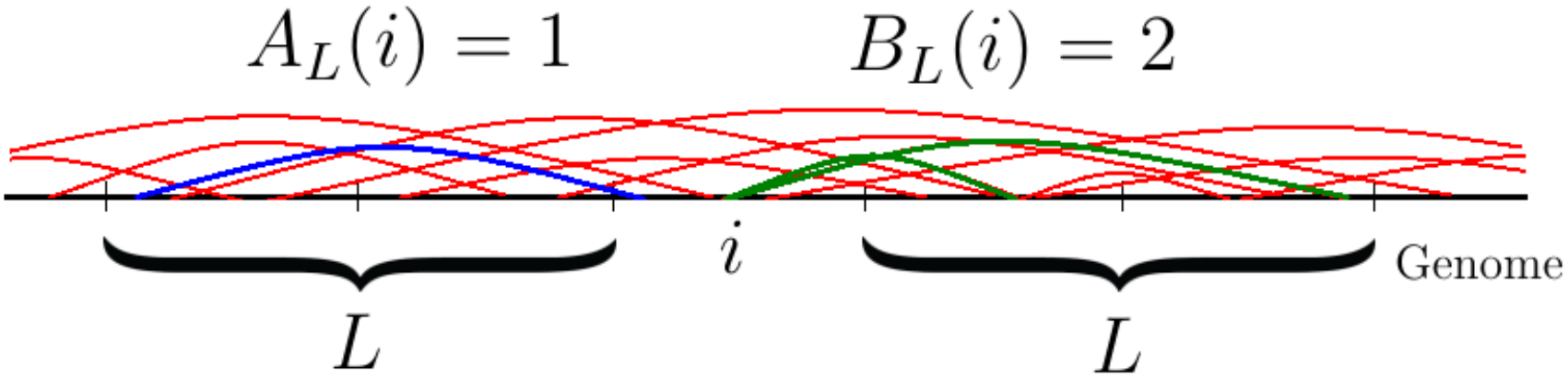
- Let's partition each chromosome into *r* bp bins, where *r* is a contact matrix resolution.
- Contacts within the chromosome can then be visualized like this. Each arc denotes a pair of reads.

Directionality index



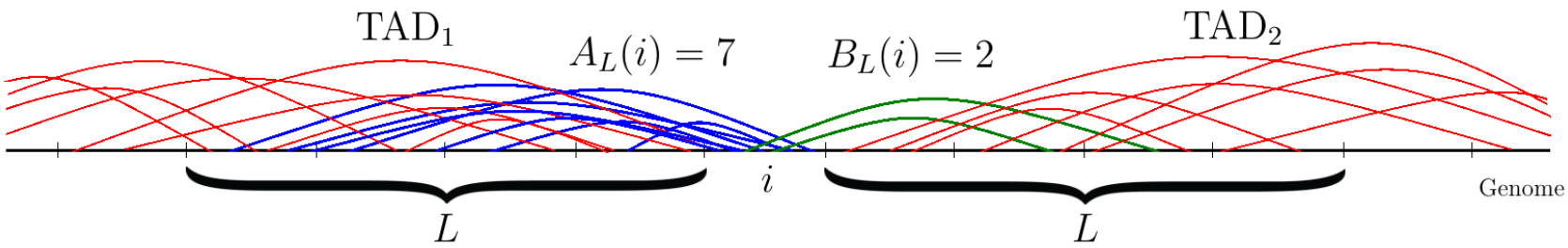
- Let's partition each chromosome into r bp bins, where r is a contact matrix resolution.
- Contacts within the chromosome can then be visualized like this. Each arc denotes a pair of reads.
- Then $A_L(i)$ is the number of read pairs that map from the bin i to the upstream L bp. L should be a multiple of r.

Directionality index



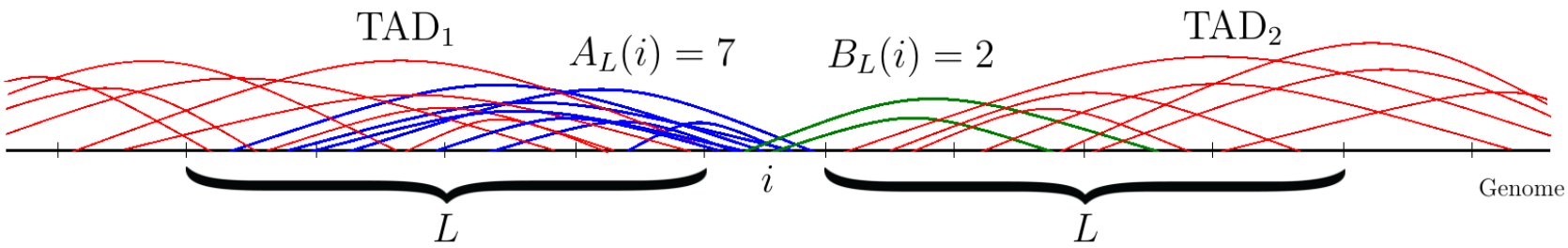
- Let's partition each chromosome into r bp bins, where r is a contact matrix resolution.
- Contacts within the chromosome can then be visualized like this. Each arc denotes a pair of reads.
- Then $A_L(i)$ is a number of read pairs that map from the bin i to the upstream L bp.
- And $B_L(i)$ is a number of read pairs that map from the bin i to the downstream L bp.

Directionality index



- At the end of a TAD we expect a bias in contact frequency towards upstream regions.

Directionality index



- At the end of a TAD we expect a bias in contact frequency towards upstream regions.
- And vice versa: at the beginning of a TAD we expect a bias in contact frequency towards downstream regions.

Directionality index

- We can use this bias for TAD calling. Consider some bin i and its L bp vicinity. Let $A \equiv A_L(i)$, $B \equiv B_L(i)$, $D \equiv D_L(i)$, and $E \equiv E_L(i)$. Then, let's define **directionality index** (Dixon et al., 2012)

$$DI = \frac{B - A}{|B - A|}\left(\frac{(A - E)^2}{E} + \frac{(B - E)^2}{E}\right),$$

where $E \equiv E_L(i) = \frac{A_L(i)+B_L(i)}{2}$ is an expected number of reads (without the upstream or downstream contact frequency bias).

Directionality index

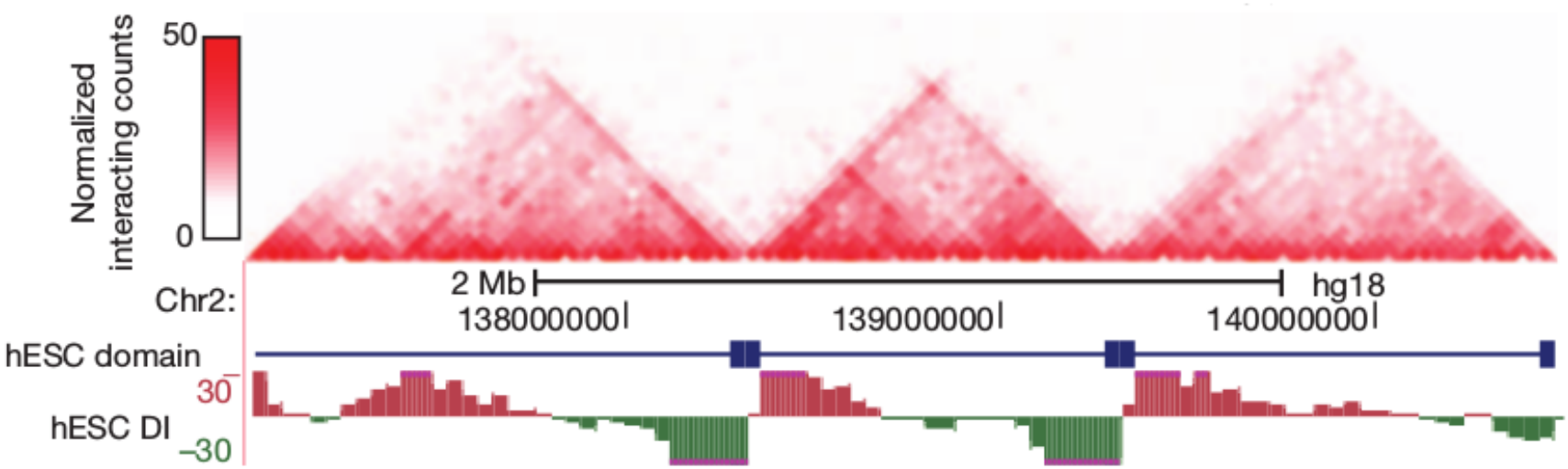
- We can use this bias for TAD calling. Consider some bin i and its L bp vicinity. Let $A \equiv A_L(i)$, $B \equiv B_L(i)$, $D \equiv D_L(i)$, and $E \equiv E_L(i)$. Then, let's define **directionality index** (Dixon et al., 2012)

$$DI = \frac{B - A}{|B - A|}\left(\frac{(A - E)^2}{E} + \frac{(B - E)^2}{E}\right),$$

  where $E \equiv E_L(i) = \frac{A_L(i)+B_L(i)}{2}$ is an expected number of reads (without the upstream or downstream contact frequency bias).
- At the end of a TAD DI should have a local minimum, and immediately at the beginning of the next TAD DI should have a local maximum.

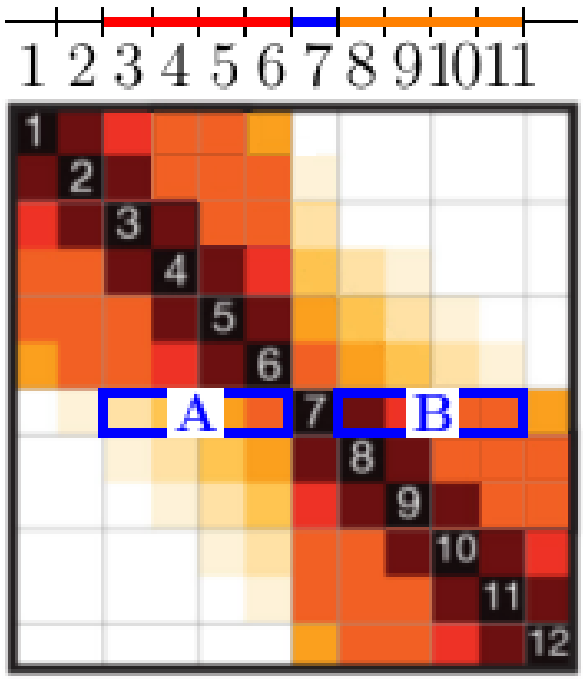
Directionality index

An illustration of this idea from Dixon et al., 2012 (Hi-C data for hESC – human embryonic stem cell line, some region of chr2):



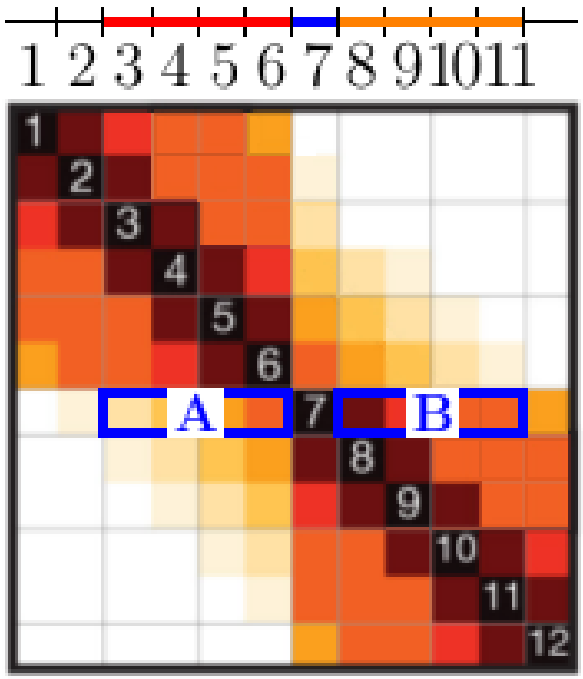
Frame Title

DI calculation from a contact matrix (fig. is based on Crane et al., 2015):



Frame Title

DI calculation from a contact matrix (fig. is based on Crane et al., 2015):

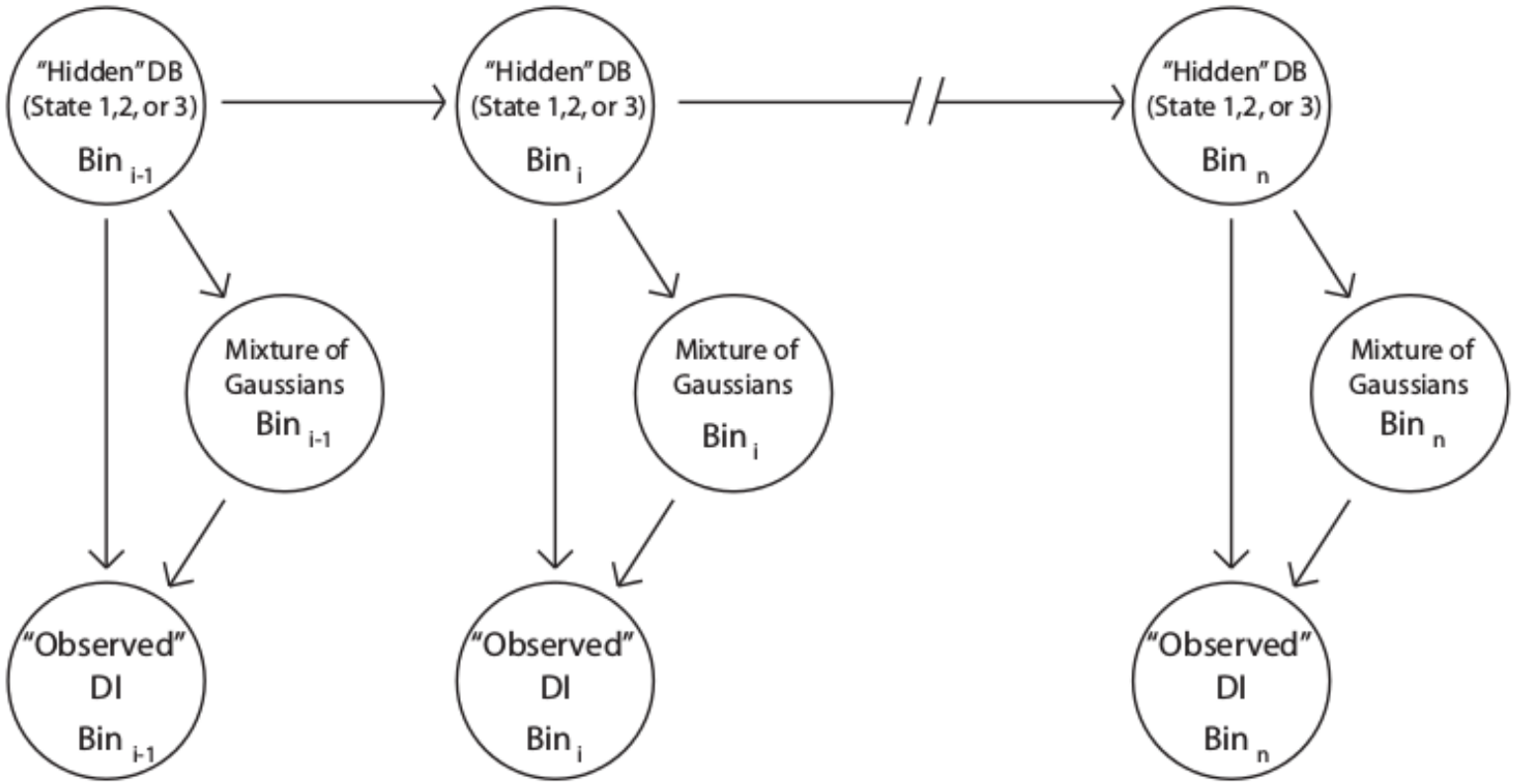


$$\mathrm{DI} = \frac{\sum_{\mathbf{B}} - \sum_{\mathbf{A}}}{|\sum_{\mathbf{B}} - \sum_{\mathbf{A}}|}\left(\frac{(\sum_{\mathbf{A}} - E)^2}{E} + \frac{(\sum_{\mathbf{B}} - E)^2}{E}\right),$$

where $E = \frac{\sum_{\mathbf{A}} + \sum_{\mathbf{B}}}{2}$, $\sum_{\mathbf{A}}$ and $\sum_{\mathbf{B}}$ are sums of elements in contact submatrices **A** and **B**, respectively.

Directionality index

Now we can define a Hidden Markov Model (HMM) for TAD calling with DI (Dixon et al., 2012):



"Upstream Bias" - State 1
"Downstream Bias" - State 2
No Bias - State 3

Directionality index

- Baum-Welch algorithm was used (somehow...) to compute maximum likelihood estimates of the model and the parameter estimates of transition and emission.

Directionality index

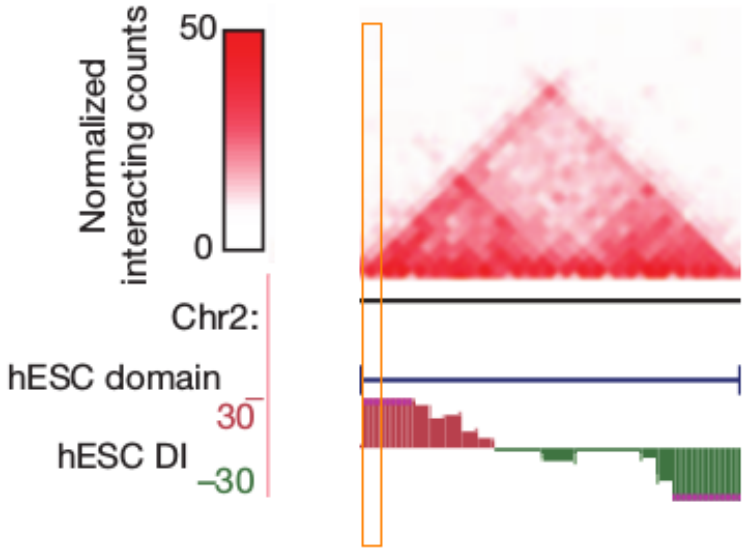
- Baum-Welch algorithm was used (somehow...) to compute maximum likelihood estimates of the model and the parameter estimates of transition and emission.
- Forward-backward algorithm was used to estimate posterior marginals, i. e., $\Pr(Q_t = q \mid D_1 = d_1, D_2 = d_2, \ldots, D_n = d_n)$, where q is a hidden state, $t \in \{1, \ldots, n\}$, $d_1, d_2, \ldots, d_n$ are emission values.

Directionality index

- Baum-Welch algorithm was used (somehow...) to compute maximum likelihood estimates of the model and the parameter estimates of transition and emission.
- Forward-backward algorithm was used to estimate posterior marginals, i. e., $\Pr(Q_t = q \mid D_1 = d_1, D_2 = d_2, \ldots, D_n = d_n)$, where q is a hidden state, $t \in \{1, \ldots, n\}$, $d_1, d_2, \ldots, d_n$ are emission values.
- For each chromosome the authors tried to use $1 - 20$ mixtures of Gaussians and chose one set with the best goodness of fit using the AIC criterion: $\mathrm{AIC} = 2k - 2\ln(L)$, where k is the number of parameters in the model and L is the maximum likelihood estimate.

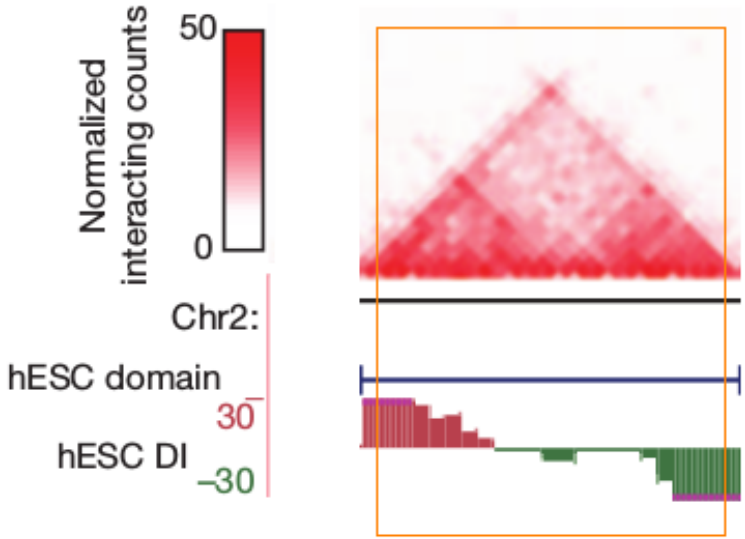
Directionality index

- **TAD calling:**
  - TAD begins at the beginning of the first `DB` state in a series of `DB` states.



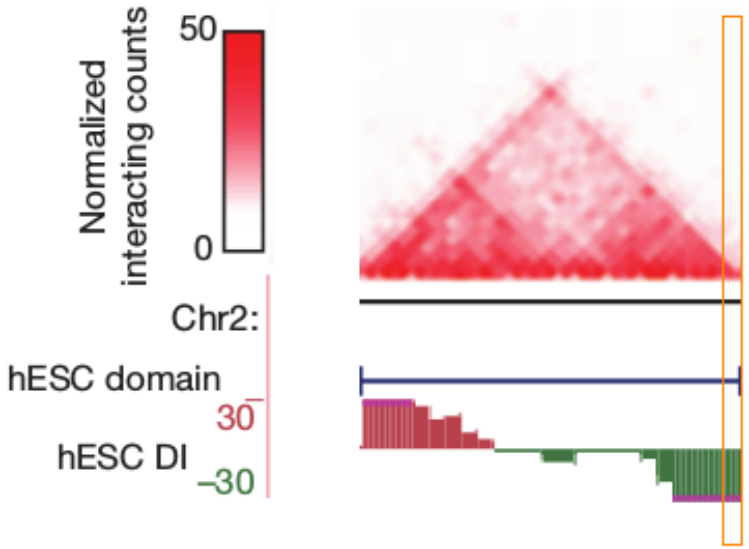
Directionality index

- **TAD calling:**
  - TAD begins at the beginning of the first `DB` state in a series of `DB` states.
  - TAD is continuous through all `DB` states in the series and then – through all the states in a `UB` series.



Directionality index

- **TAD calling:**
  - TAD begins at the beginning of the first `DB` state in a series of `DB` states.
  - TAD is continuous through all `DB` states in the series and then – through all the states in a `UB` series.
  - TAD ends in the last `UB` state in the series of `UB` states.



Directionality index

- **TAD calling:**
  - TAD begins at the beginning of the first `DB` state in a series of `DB` states.
  - TAD is continuous through all `DB` states in the series and then – through all the states in a `UB` series.
  - TAD ends at the end of the last `UB` state in the series of `UB` states.
- **TAD borders:** a region between TADs is called **topological boundary** if its length is less than 400 kbp, otherwise it is called **unrecognized chromatin**.

Directionality index

- **TAD calling:**
  - TAD begins at the beginning of the first `DB` state in a series of `DB` states.
  - TAD is continuous through all `DB` states in the series and then – through all the states in a `UB` series.
  - TAD ends at the end of the last `UB` state in the series of `UB` states.
- **TAD borders:** a region between TADs is called **topological boundary** if its length is less than 400 kbp, otherwise it is called **unrecognized chromatin**.
- Topological boundaries in mouse ESC were found to be quite small, 76.33 % of them being less than 50 kbp.

Directionality index

The main biological results in Dixon et al., 2012 are as follows:

- TADs were called in mouse and human ESC, as well as in some terminally differentiated cell types. E. g., about 91 % of the mouse ESC is occupied by TADs with median size around 880 kbp.

Directionality index

The main biological results in Dixon et al., 2012 are as follows:

- TADs were called in mouse and human ESC, as well as in some terminally differentiated cell types. E. g., about 91 % of the mouse ESC is occupied by TADs with median size around 880 kbp.
- TADs are stable across different cell types and highly conserved across species.

Directionality index

The main biological results in Dixon et al., 2012 are as follows:

- TADs were called in mouse and human ESC, as well as in some terminally differentiated cell types. E. g., about 91 % of the mouse ESC is occupied by TADs with median size around 880 kbp.
- TADs are stable across different cell types and highly conserved across species.
- TAD borders are enriched for CTCF, housekeeping genes, tRNAs, and SINE retrotransposons.

Directionality index

The main biological results in Dixon et al., 2012 are as follows:

- TADs were called in mouse and human ESC, as well as in some terminally differentiated cell types. E. g., about 91 % of the mouse ESC is occupied by TADs with median size around 880 kbp.
- TADs are stable across different cell types and highly conserved across species.
- TAD borders are enriched for CTCF, housekeeping genes, tRNAs, and SINE retrotransposons.

These results (and raw Hi-C data from the paper) are used in biological studies (see, e. g., Battulin et al., 2015, Rao et al., 2014, Van Bortle, 2014, Pope et al, 2014, Duggal et al., 2014, Kolovos et al., 2014, Zhao et al., 2013, Lu et al, 2013)

Directionality index

The main biological results in Dixon et al., 2012 are as follows:

- TADs were called in mouse and human ESC, as well as in some terminally differentiated cell types. E. g., about 91 % of the mouse ESC is occupied by TADs with median size around 880 kbp.
- TADs are stable across different cell types and highly conserved across species.
- TAD borders are enriched for CTCF, housekeeping genes, tRNAs, and SINE retrotransposons.

These results (and raw Hi-C data from the paper) are used in biological studies (see, e. g., Battulin et al., 2015, Rao et al., 2014, Van Bortle, 2014, Pope et al, 2014, Duggal et al., 2014, Kolovos et al., 2014, Zhao et al., 2013, Lu et al, 2013), as well as in papers on Hi-C processing tools and methods (see Roy et al, 2015, Weinreb et al., 2015, Filippova el al., 2014, Rao et al., 2014, Shavit et al., 2014, Lu et al, 2013, Merelli et al., 2013).

Directionality index

Although Dixon et al., 2012 didn't publish their scripts (they used MATLAB) and detailed description of the HMM, directionality index (DI) became a popular metric for TAD calling.

Directionality index

Although Dixon et al., 2012 didn't publish their MATLAB scripts and detailed description of the HMM, directionality index (DI) became a popular metric for TAD calling. E. g.:

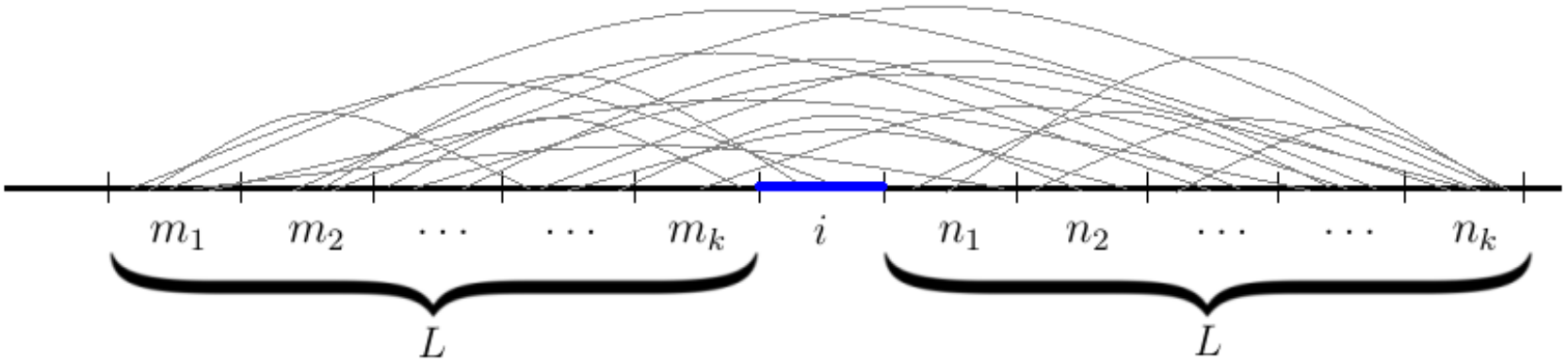
- Pope et al, 2014 called TAD borders (without HMM) in human fibroblasts IMR90 in order to compare them to those previously called in Dixon et al., 2012 (higher resolution Hi-C data were used) and to use them in replication-timing studies.

Directionality index

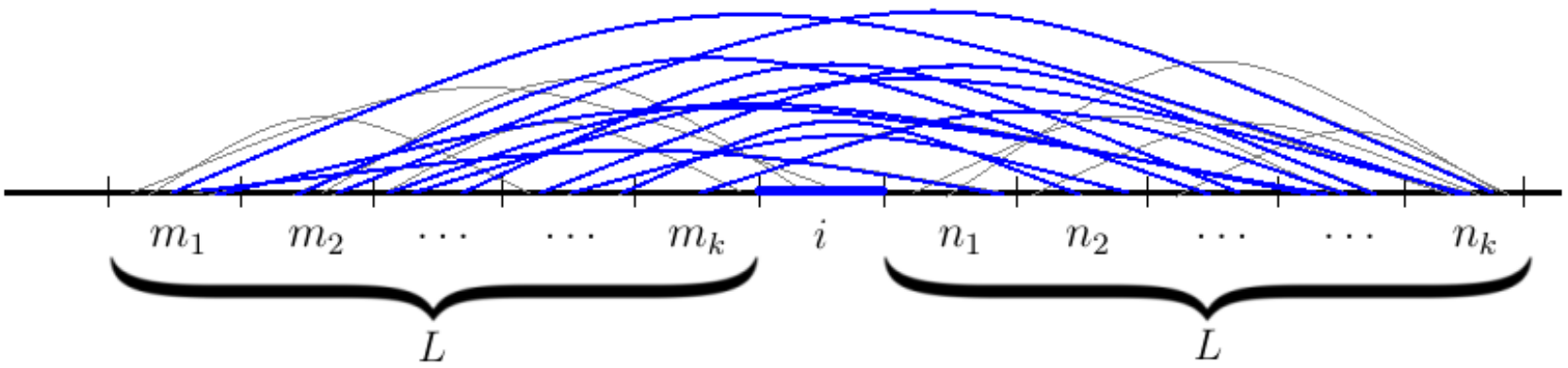
Although Dixon et al., 2012 didn't publish their MATLAB scripts and detailed description of the HMM, directionality index (DI) became a popular metric for TAD calling. E. g.:

- Pope et al, 2014 called TAD borders (without HMM) in human fibroblasts IMR90 in order to compare them to those previously called in Dixon et al., 2012 (higher resolution Hi-C data were used) and to use them in replication-timing studies.
- Dileep et al., 2015 calculated DI in six regions at several time points in the G1-phase of mouse mammary epithelial cell line (C127) watching a switch from a negligible to strong directionality bias that suggested formation of TADs.

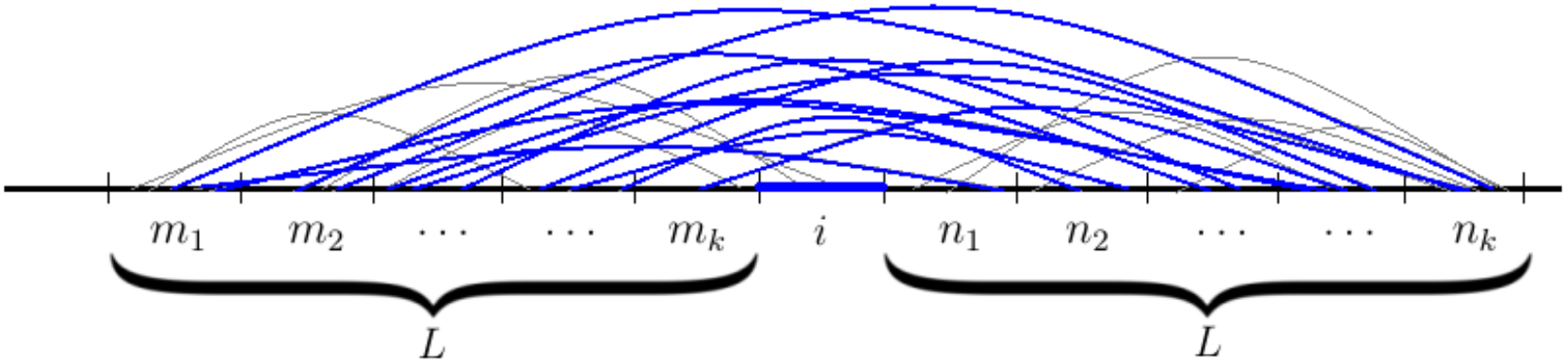
Insulation score



Insulation score



Insulation score

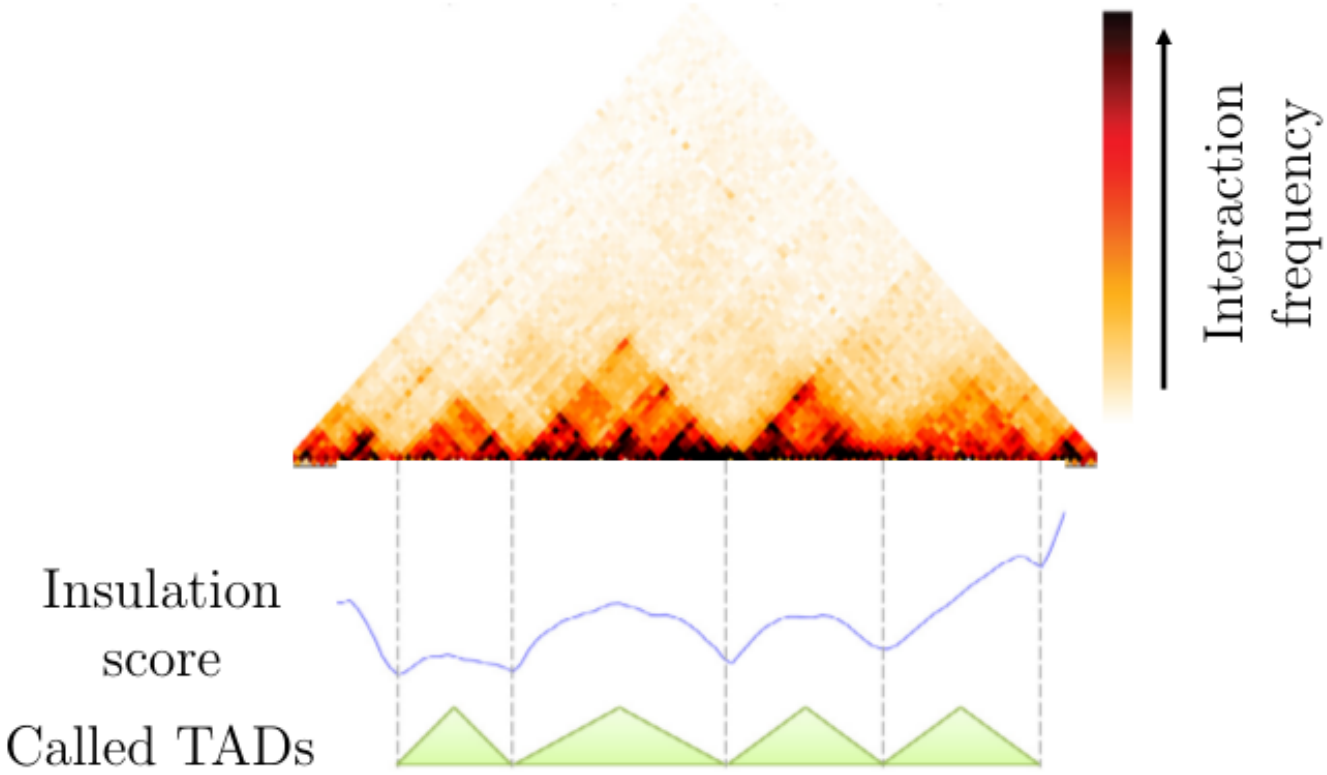


Insulation score (IS) is defined for a bin as an average number of interactions that occur across this bin in some vicinity of the bin (Crane et al., 2015):

$$\text{IS} = \frac{1}{k^2} \sum_{m \in M,\, n \in N} C(m, n),$$

where $N = \{n_1, n_2, \ldots, n_k\}$, $M = \{m_1, m_2, \ldots, m_k\}$, $C(m, n)$ is a number of interactions between bin m and bin n.

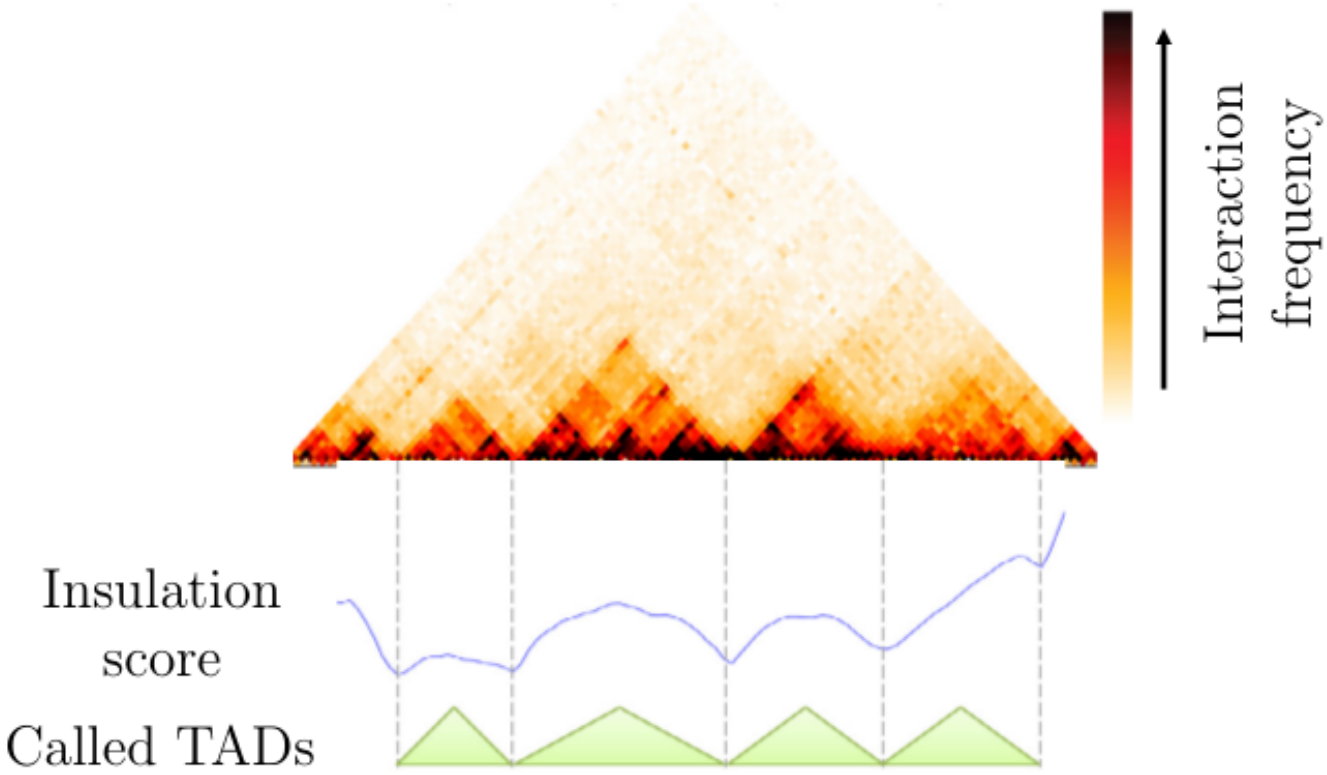
Insulation score



Lajoie et al., 2015, adapted

- We expect that IS has local minimums at TAD borders.

Insulation score

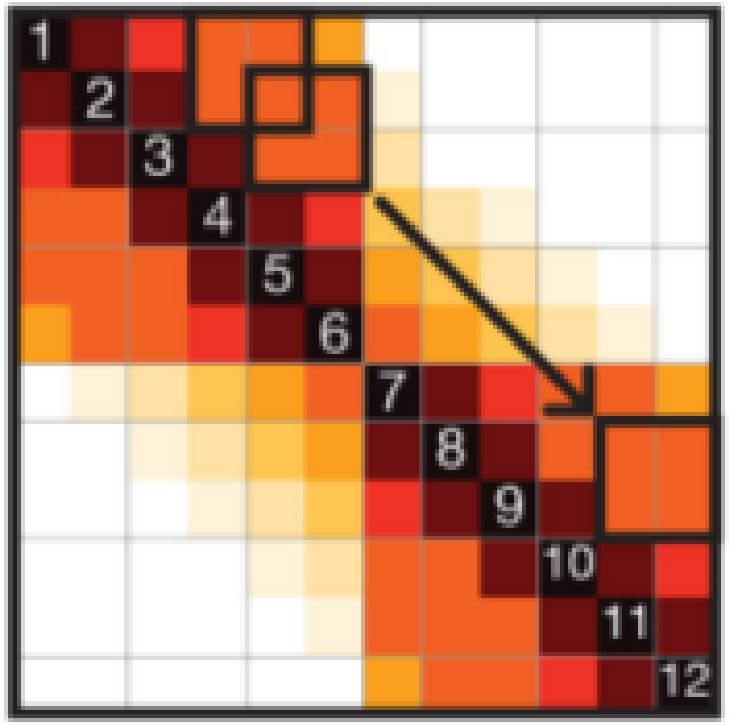


Lajoie et al., 2015, adapted

- We expect that IS has local minimums at TAD borders.
- IS plot is often called **insulation profile**.

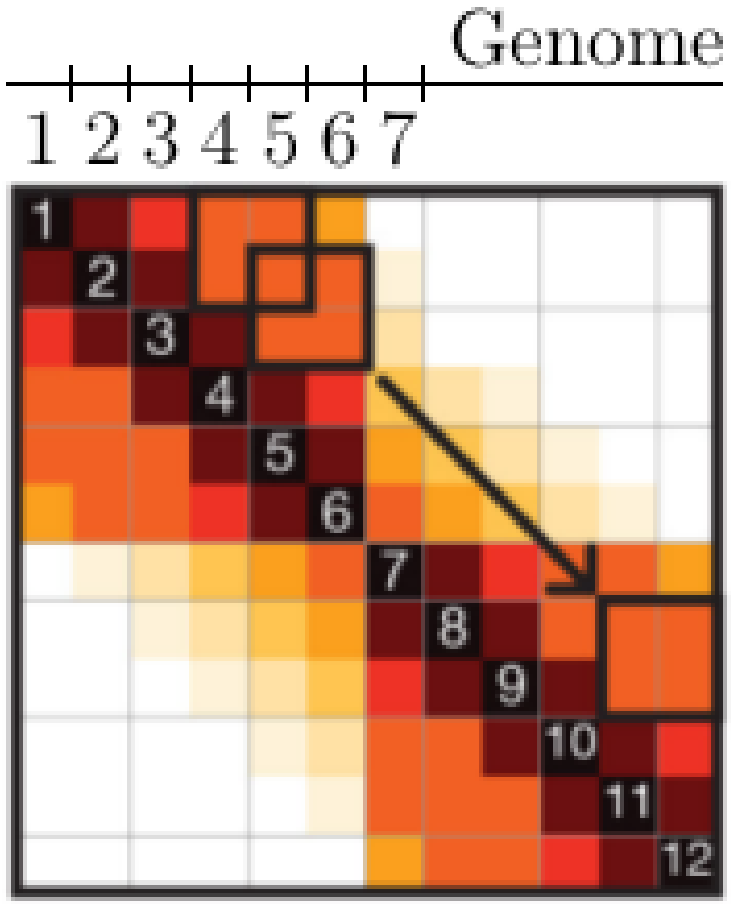
Insulation score

IS can be calculated using a square window sliding along the diagonal of a contact matrix: average number of interactions in this window is the insulation score value (Crane et al., 2015, adapted):



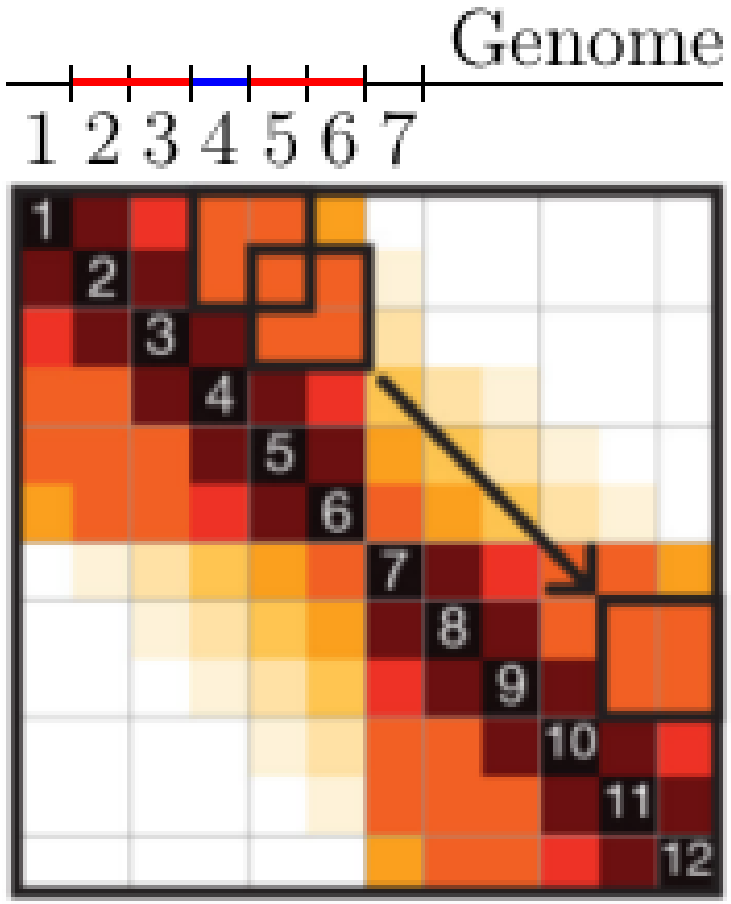
Insulation score

IS can be calculated using a square window sliding along the diagonal of a contact matrix: average number of interactions in this window is the insulation score value (Crane et al., 2015, adapted):



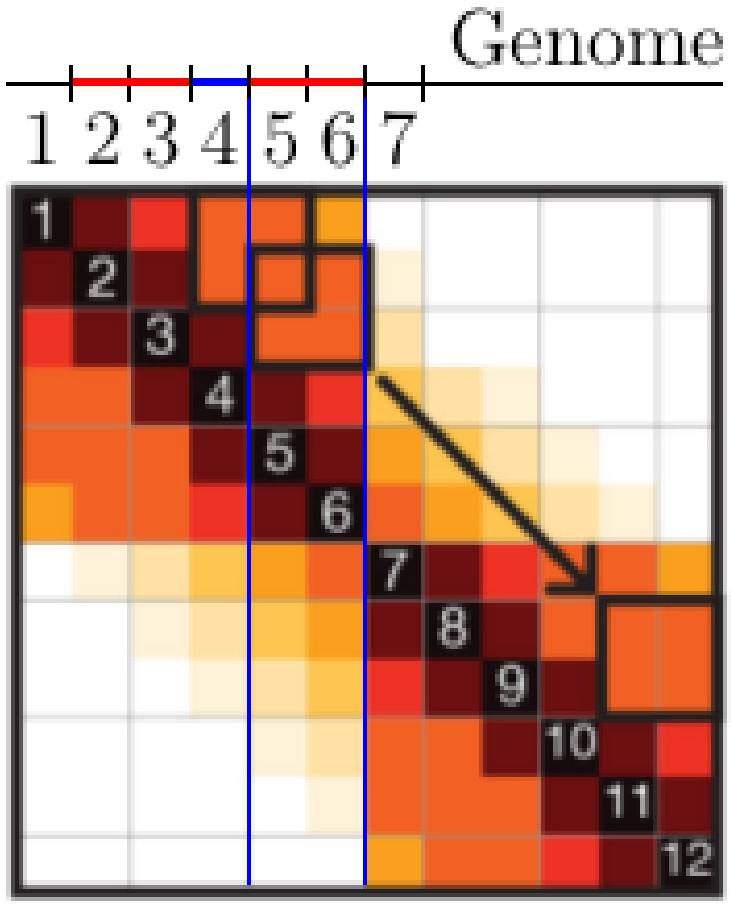
Insulation score

IS can be calculated using a square window sliding along the diagonal of a contact matrix: average number of interactions in this window is the insulation score value (Crane et al., 2015, adapted):



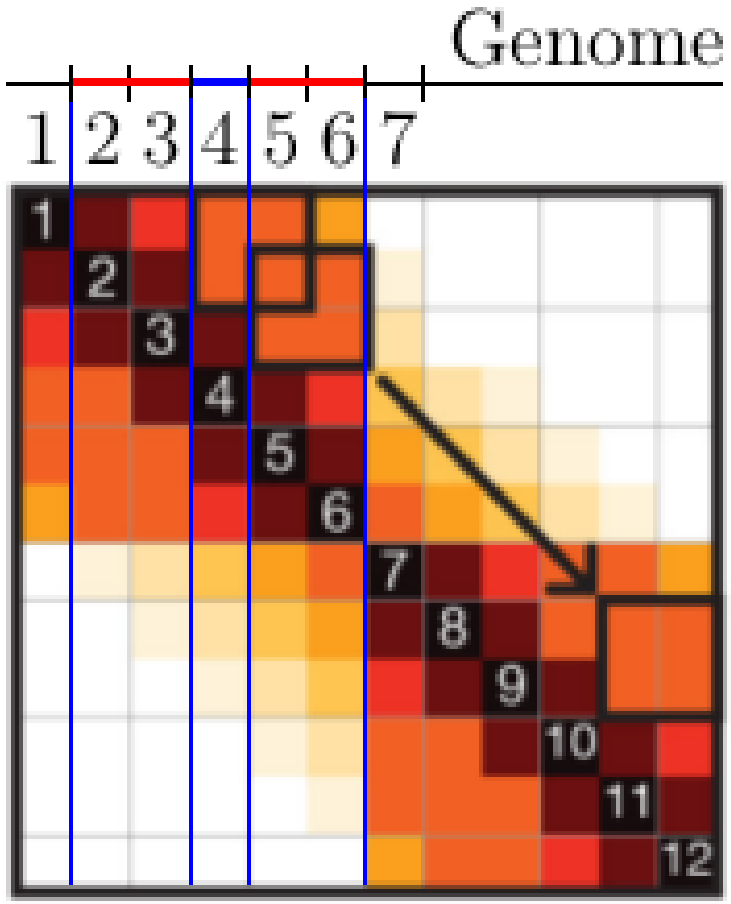
Insulation score

IS can be calculated using a square window sliding along the diagonal of a contact matrix: average number of interactions in this window is the insulation score value (Crane et al., 2015, adapted):



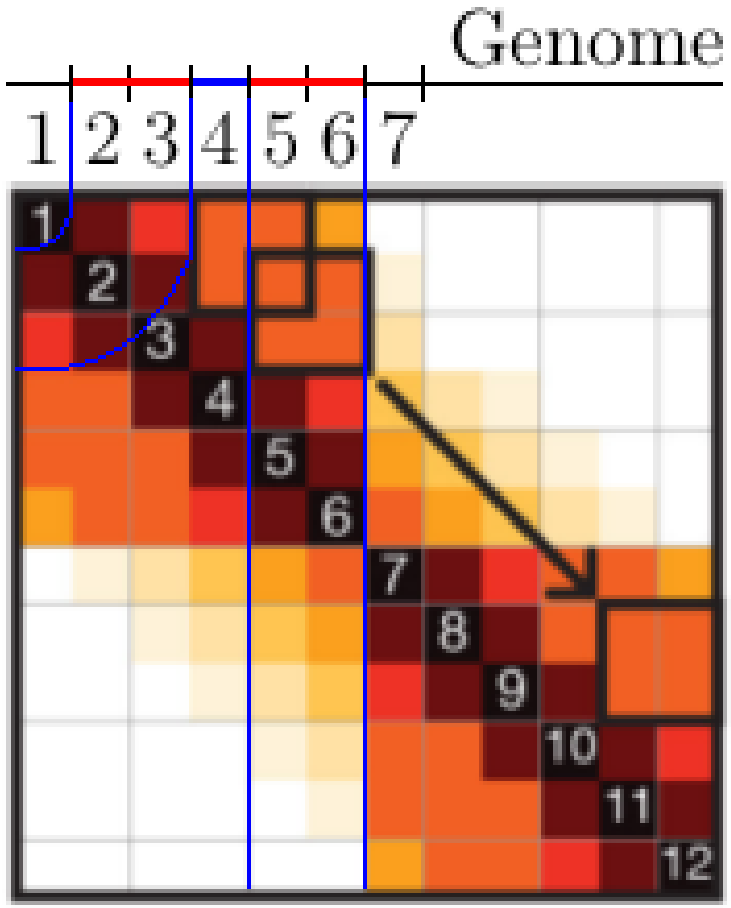
Insulation score

IS can be calculated using a square window sliding along the diagonal of a contact matrix: average number of interactions in this window is the insulation score value (Crane et al., 2015, adapted):



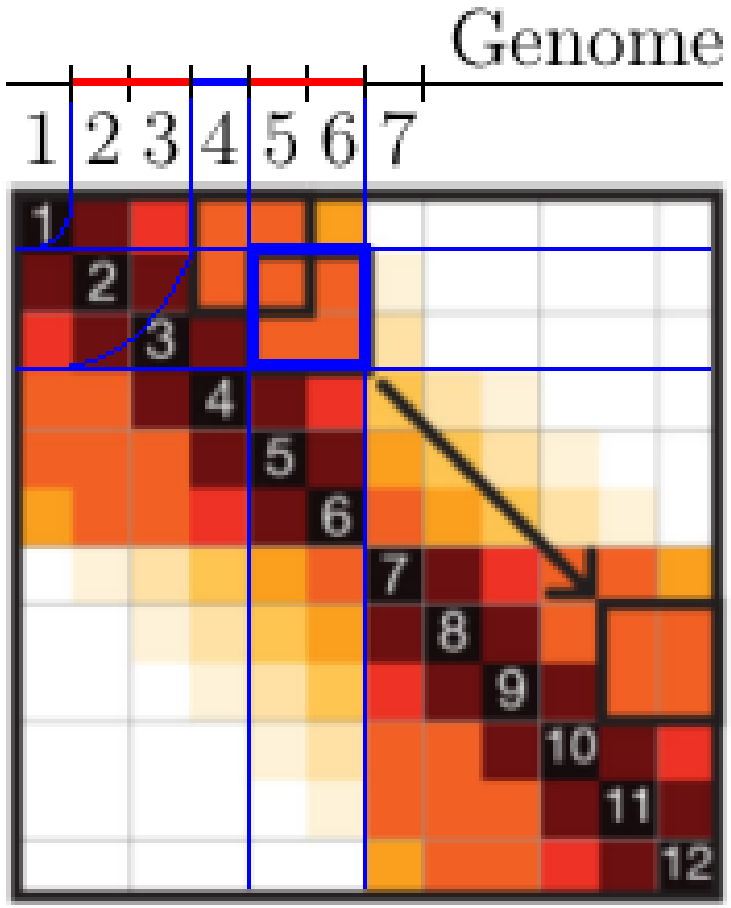
Insulation score

IS can be calculated using a square window sliding along the diagonal of a contact matrix: average number of interactions in this window is the insulation score value (Crane et al., 2015, adapted):



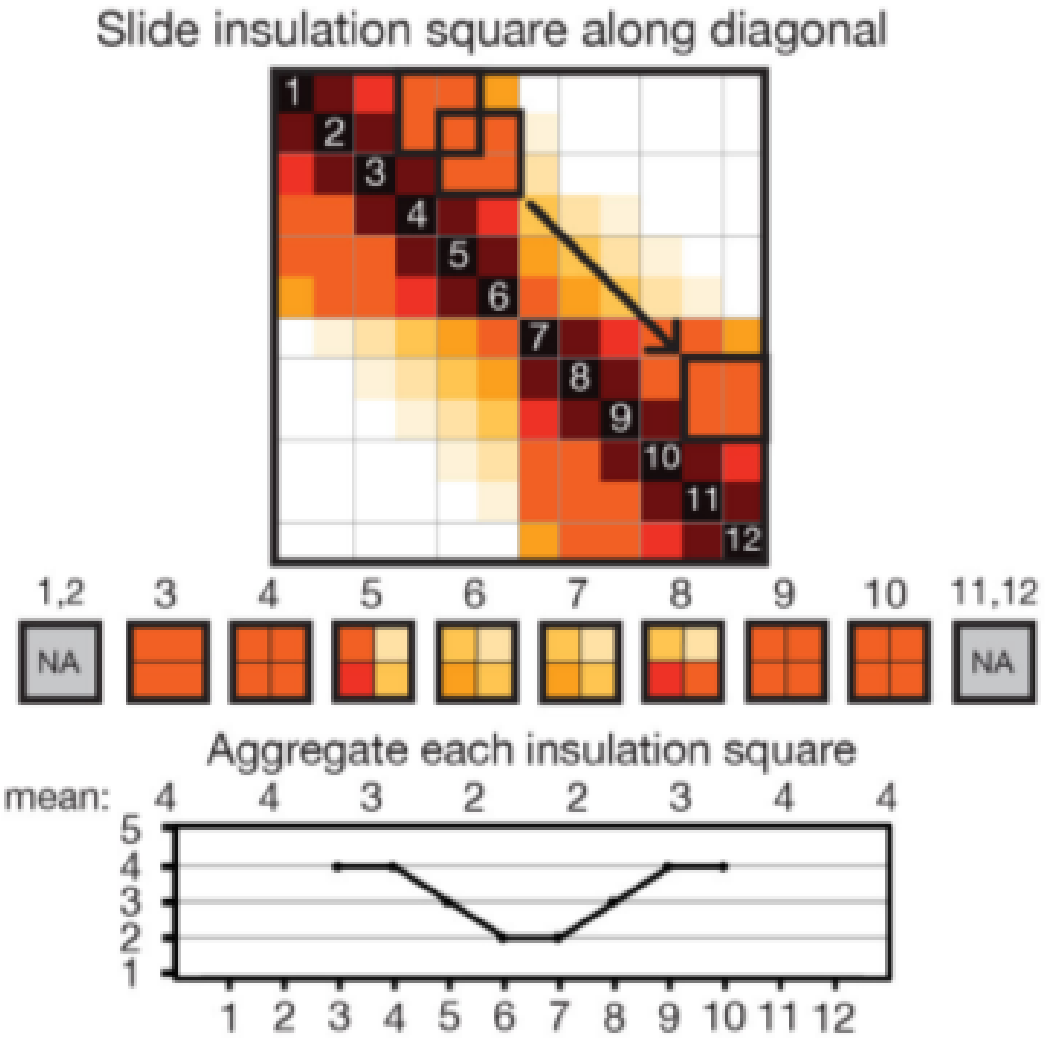
Insulation score

IS can be calculated using a square window sliding along the diagonal of a contact matrix: average number of interactions in this window is the insulation score value (Crane et al., 2015, adapted):



Insulation score

IS calculation scheme (Crane et al., 2015):



Insulation score

TAD calling with IS (Crane et al., 2015):

- Calculate IS along a chromosome.

Insulation score

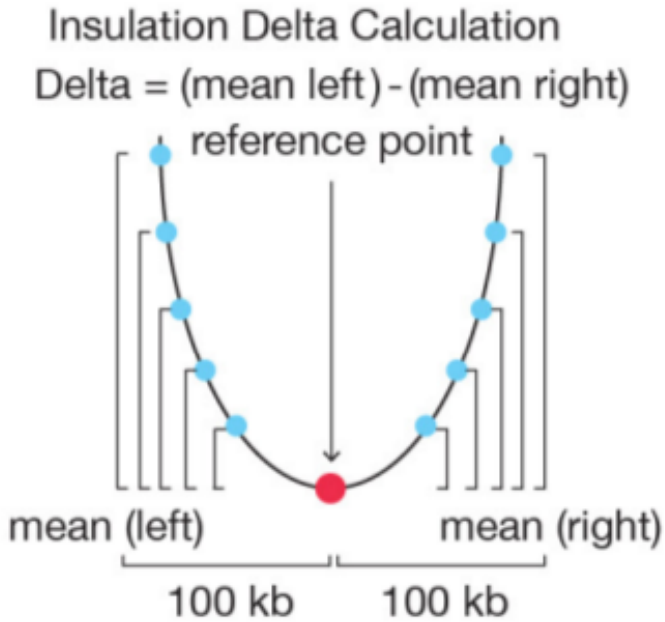
TAD calling with IS (Crane et al., 2015):

- Calculate IS along a chromosome.
- Normalize each IS value: $\mathrm{IS} := \log_2 \frac{\mathrm{IS}}{\mathrm{IS}_{\mathrm{avg}}}$, where $\mathrm{IS}_{\mathrm{avg}}$ is the mean of all IS values for the chromosome.

Insulation score

TAD calling with IS (Crane et al., 2015):

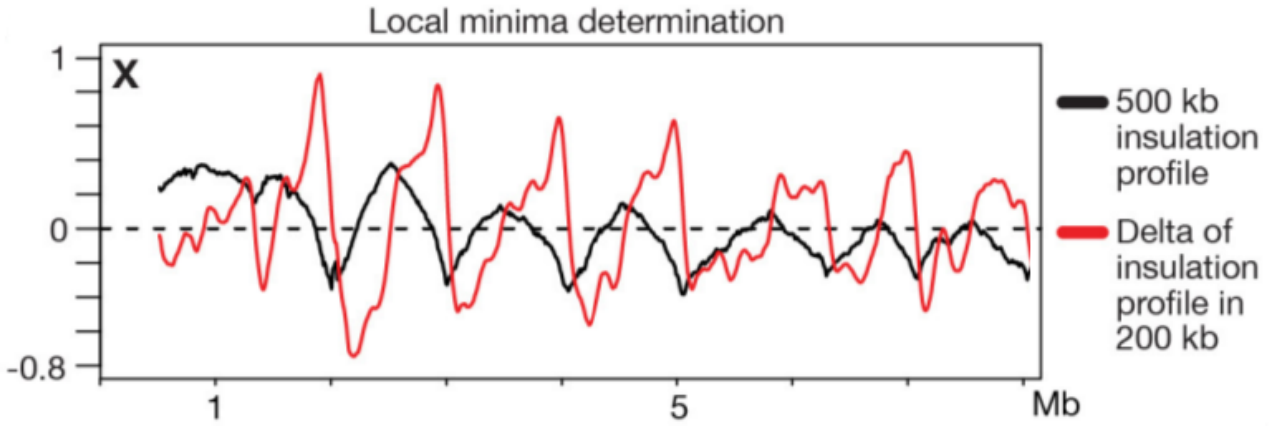
- Calculate IS along a chromosome.
- Normalize each IS value: $\text{IS} := \log_2 \frac{\text{IS}}{\text{IS}_{\text{avg}}}$, where IS_{avg} is the mean of all IS values for the chromosome.
- Calculate **Δ** values for each bin *i* (Crane et al., 2015, Extended Data):



Insulation score

TAD calling with IS (Crane et al., 2015):

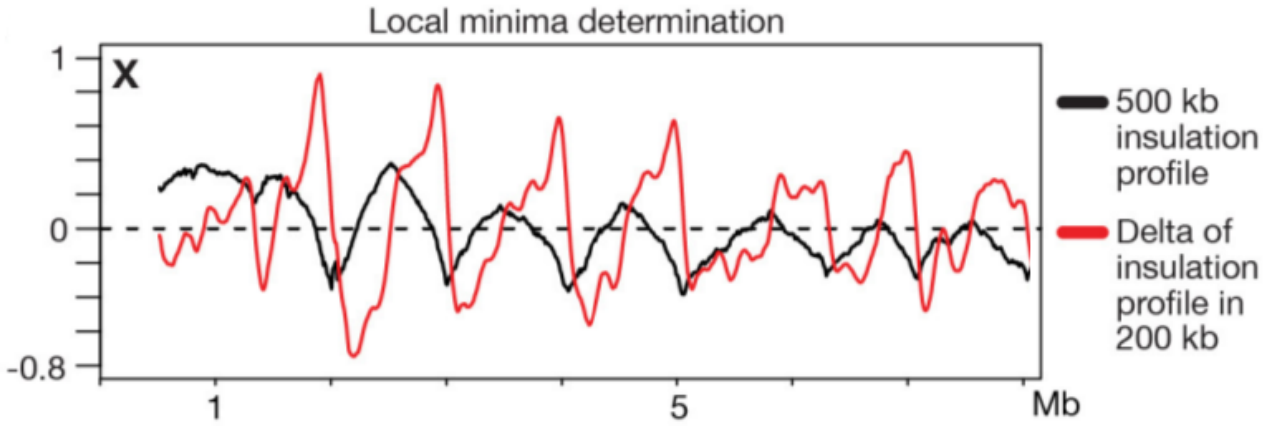
- Calculate IS along a chromosome.
- Normalize each IS value: $\text{IS} := \log_2 \frac{\text{IS}}{\text{IS}_{\text{avg}}}$, where IS_{avg} is the mean of all IS values for the chromosome.
- Calculate **Δ** values for each bin i. $\Delta_i = 0$ at all IS peaks and valleys (minimums) (Crane et al., 2015, adapted):



Insulation score

TAD calling with IS (Crane et al., 2015):

- Calculate IS along a chromosome.
- Normalize each IS value: $\mathrm{IS} := \log_2 \frac{\mathrm{IS}}{\mathrm{IS}_{\mathrm{avg}}}$, where $\mathrm{IS}_{\mathrm{avg}}$ is the mean of all IS values for the chromosome.
- Calculate **Δ** values for each bin i. $\Delta_i = 0$ at all IS peaks and valleys (minimums) (Crane et al., 2015, adapted):



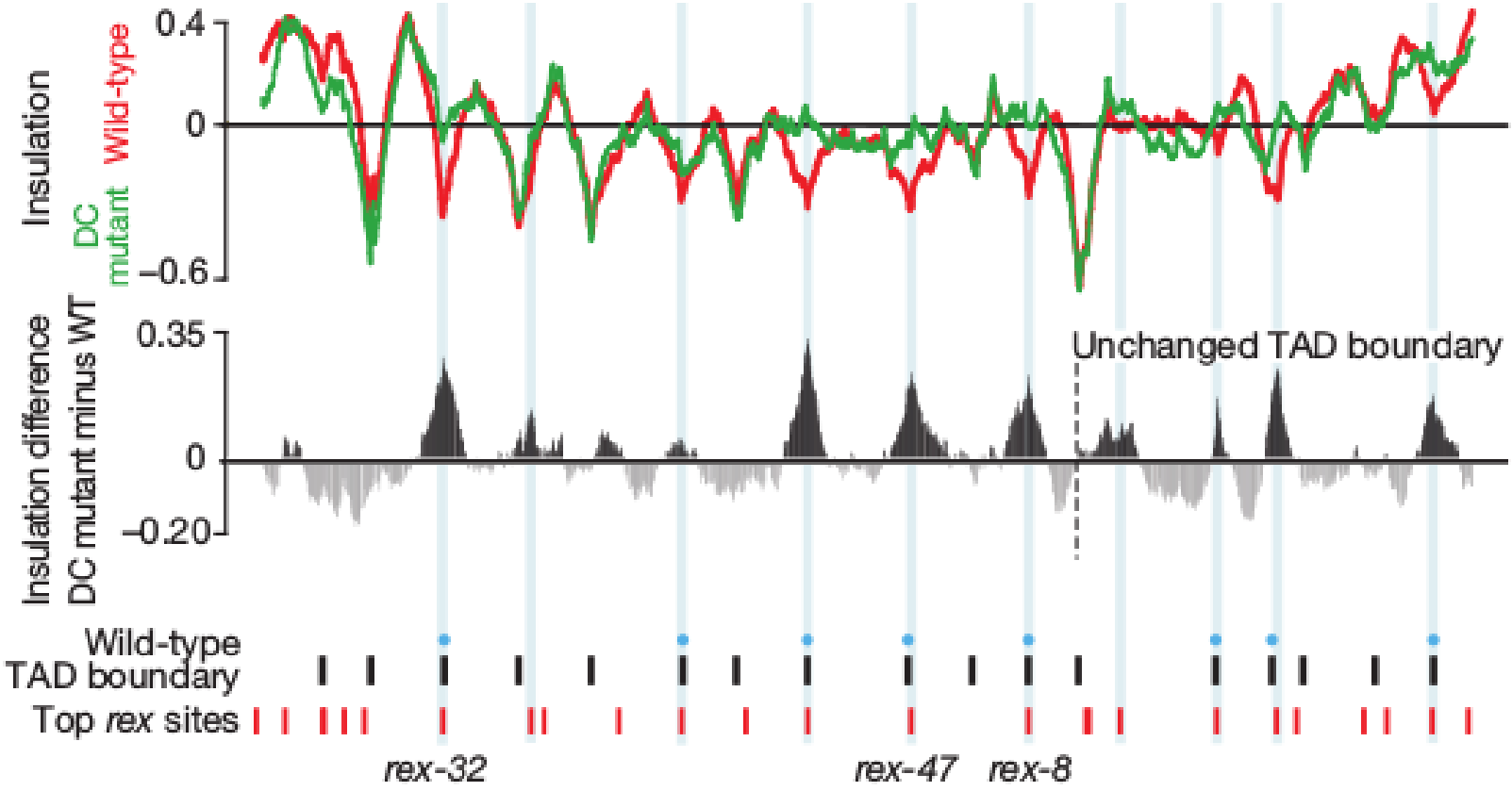
- TAD border is called at bin i if $\Delta_i = 0$, the nearest Δ local max (Δ_{max}) is to the left of bin i, the nearest Δ local min (Δ_{min}) is to the right, and $S_i \equiv \Delta_{\mathrm{max}} - \Delta_{\mathrm{min}} > 0.1$. S_i is called **border (boundary) strength**. TAD is called between two borders.

Insulation score

- Crane et al., 2015 published their Perl script for TAD calling with IS.

Insulation score

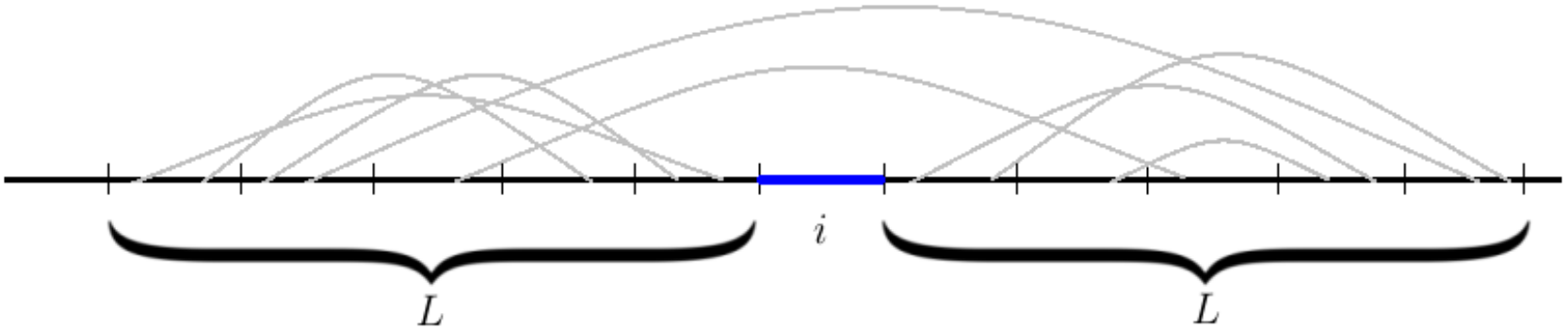
- Crane et al., 2015 published their Perl script for TAD calling with IS.
- They called TAD borders with IS to see how they change in *C. elegans* X chromosome due to dosage compensation complex (DCC) depletion (Crane et al., 2015, adapted):



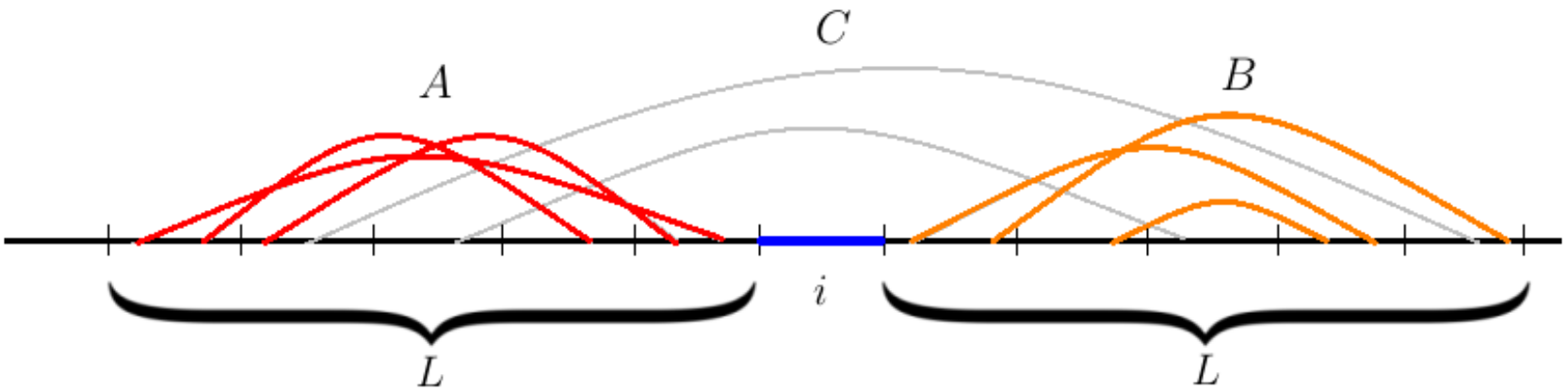
Insulation score

- Crane et al., 2015 published their Perl script for TAD calling with IS.
- They called TAD borders with IS to see how they change in *C. elegans* X chromosome due to dosage compensation complex (DCC) depletion.
- Barutcu et al., 2015 called TADs with IS to see differences in higher order chromatin structure between MCF-10A mammary epithelial and MCF-7 breast cancer cell lines.

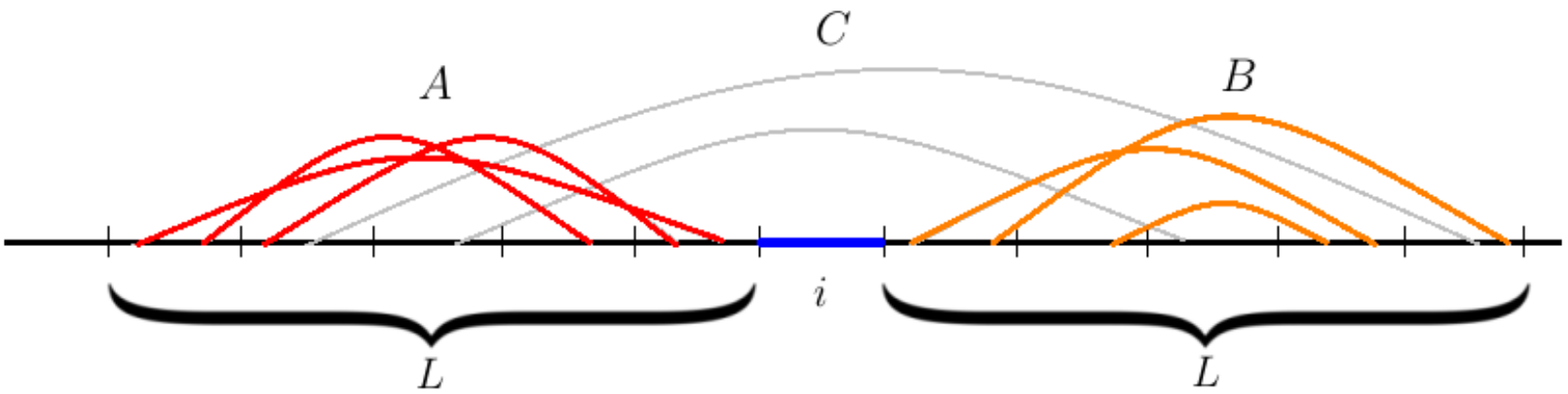
Contrast index



Contrast index



Contrast index



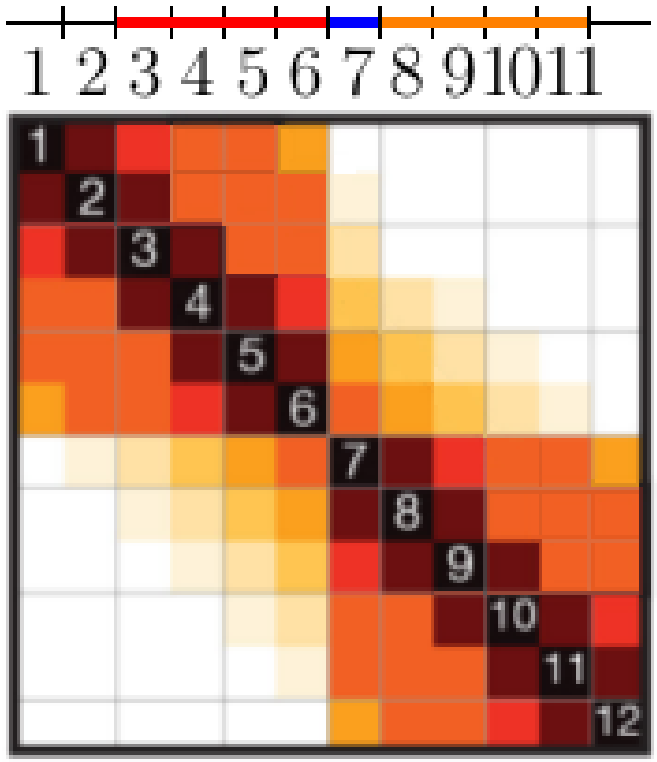
Contrast index is defined as follows (Van Bortle et al., 2014, Alekseyenko et al., 2015):

$$\mathrm{CI} = \frac{A + B}{C},$$

where A is a total number of interactions to the left of bin i in L-vicinity, B is a total number of interactions to the right of bin i in L-vicinity, and C is a number of interactions that occur over bin i from the left L-vicinity to the right.

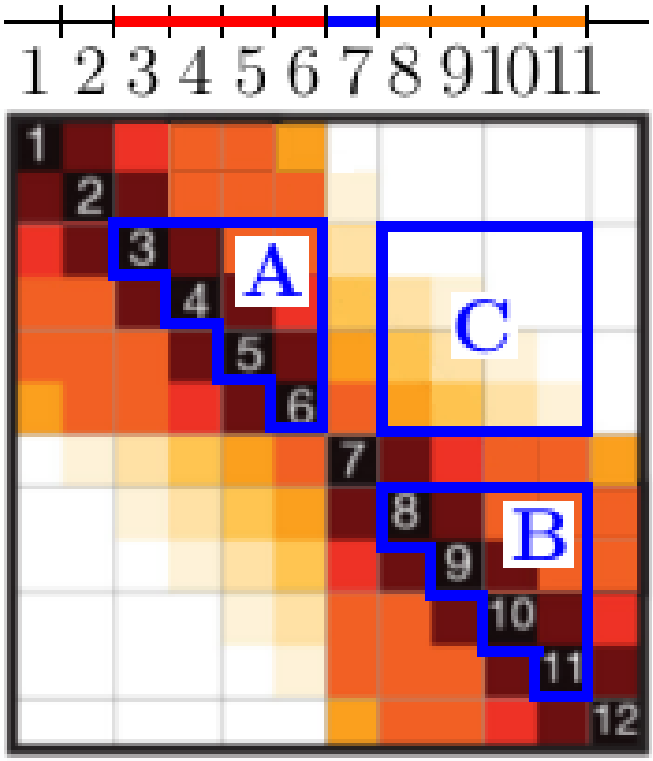
Contrast index

CI calculation using a contact matrix (fig. is based on Crane et al., 2015):



Contrast index

CI calculation using a contact matrix (fig. is based on Crane et al., 2015):



$$\mathrm{CI} = \frac{\sum_{\mathbf{A}} + \sum_{\mathbf{B}}}{\sum_{\mathbf{C}}},$$

where $\sum_{\mathbf{A}}$, $\sum_{\mathbf{B}}$, $\sum_{\mathbf{C}}$ are sums of elements in **A**, **B**, and **C** contact submatrices, respectively.

Contrast index

- TAD is called between two bins with CI values higher than some threshold.

Contrast index

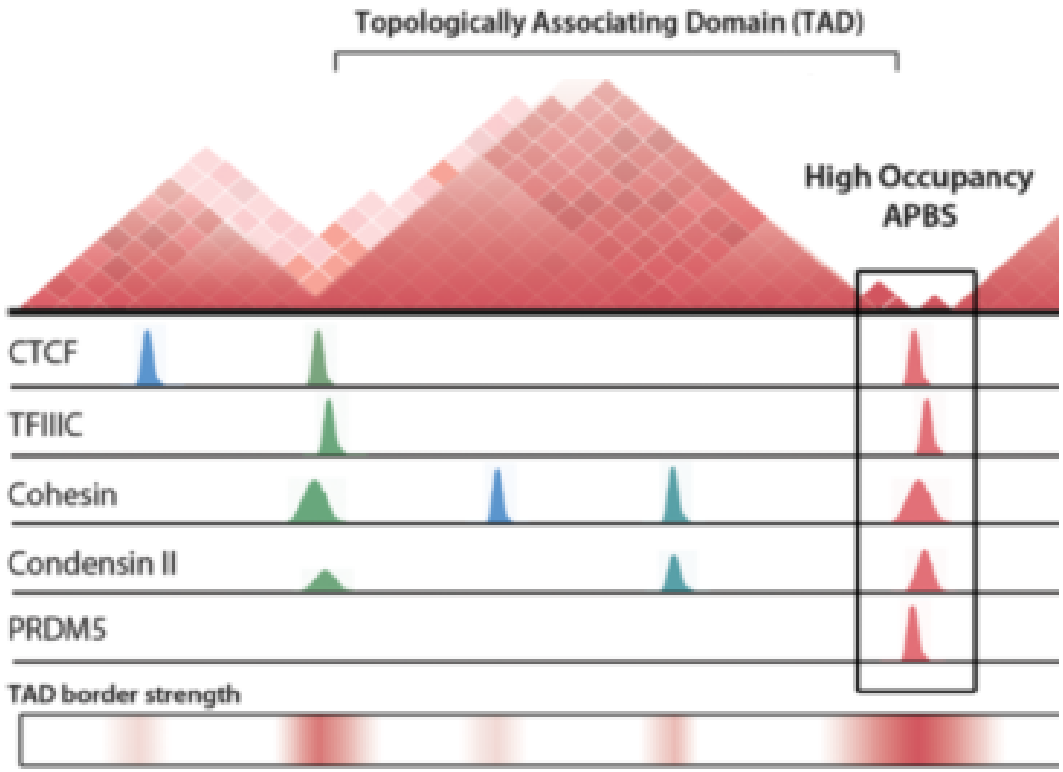
- TAD is called between two bins with CI values higher than some threshold.
- No tool (script) was published for CI calculation.

Contrast index

- TAD is called between two bins with CI values higher than some threshold.
- No tool (script) was published for CI calculation.
- CI was used for TAD calling and TAD border strength assessment in several papers.

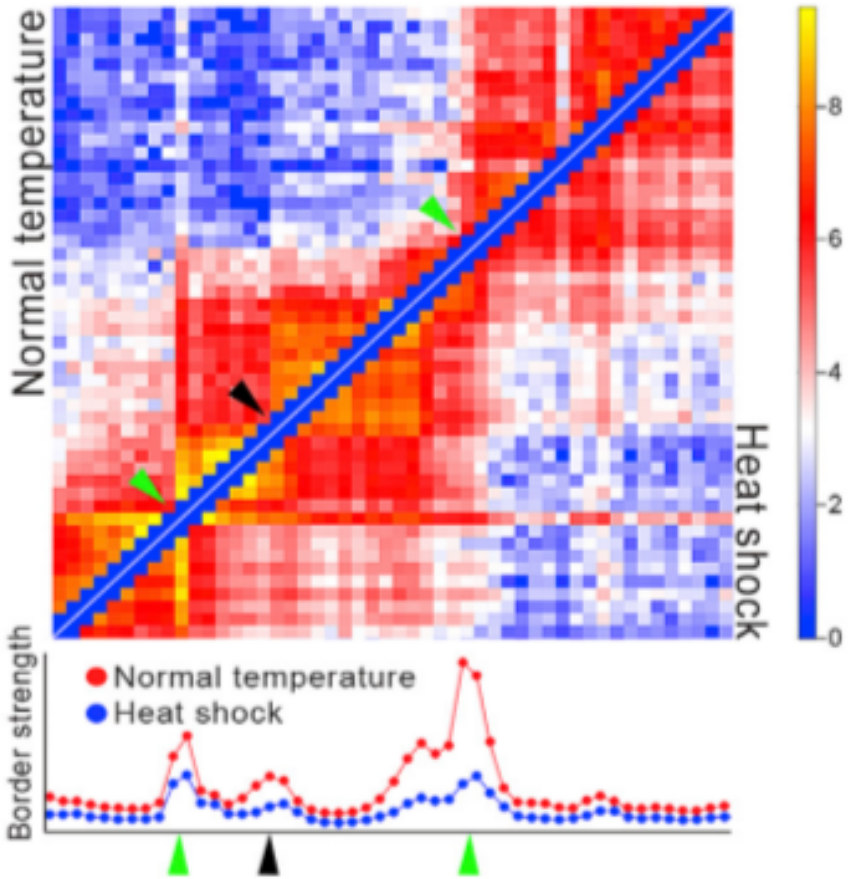
Contrast index

- CI was used for TAD calling and TAD border strength assessment in several papers. E. g.:
  - Van Bortle et al., 2014 studied a relationship between TAD border strength and architectural proteins binding site (APBS) abundance (fig. is adapted):



Contrast index

- CI was used for TAD calling and TAD border strength assessment in several papers. E. g.:
  - Li et al., 2015 studied TAD border strength decline in *Drosophila* cells after heat-shock:



Outline

Conclusion

- **TADs** are stable and evolutionary conserved units of transcription regulation in mammals. Some similar self-interacting domains were found in other Eukaryotic species.

Conclusion

- **TADs** are stable and evolutionary conserved units of transcription regulation in mammals. Some similar self-interacting domains were found in other Eukaryotic species.
- **Pros** and **cons** of **considered TAD calling methods**:

Conclusion

- **TADs** are stable and evolutionary conserved units of transcription regulation in mammals. Some similar self-interacting domains were found in other Eukaryotic species.
- **Pros** and **cons** of **considered TAD calling methods**:
  - DI, IS, and CI are intuitive and inferred directly from TAD definition.

Conclusion

- **TADs** are stable and evolutionary conserved units of transcription regulation in mammals. Some similar self-interacting domains were found in other Eukaryotic species.
- **Pros** and **cons** of **considered TAD calling methods**:
  - DI, IS, and CI are intuitive and inferred directly from TAD definition.
  - They can be used both for TAD calling and TAD border strength assessment.

Conclusion

- **TADs** are stable and evolutionary conserved units of transcription regulation in mammals. Some similar self-interacting domains were found in other Eukaryotic species.
- **Pros** and **cons** of **considered TAD calling methods**:
  - DI, IS, and CI are intuitive and inferred directly from TAD definition.
  - They can be used both for TAD calling and TAD border strength assessment.
  - DI, IS, and CI are easy to compute: each of them can be calculated in $O(NK)$ time for one chromosome, where N is a number of bins in a chromosome, and $2K$ is a number of bins in the $2L$-vicinity of each bin. Typically, K is much less than N.

Conclusion

- **TADs** are stable and evolutionary conserved units of transcription regulation in mammals. Some similar self-interacting domains were found in other Eukaryotic species.
- **Pros** and **cons** of **considered TAD calling methods**:
  - DI, IS, and CI are intuitive and inferred directly from TAD definition.
  - They can be used both for TAD calling and TAD border strength assessment.
  - DI, IS, and CI are easy to compute: each of them can be calculated in $O(NK)$ time for one chromosome, where N is a number of bins in a chromosome, and $2K$ is a number of bins in the $2L$-vicinity of each bin. Typically, K is much less than N.
  - We need an arbitrary threshold / percentile or a kind of HMM to call TADs with these metrics.

Conclusion

- **TADs** are stable and evolutionary conserved units of transcription regulation in mammals. Some similar self-interacting domains were found in other Eukaryotic species.
- **Pros** and **cons** of **considered TAD calling methods**:
  - DI, IS, and CI are intuitive and inferred directly from TAD definition.
  - They can be used both for TAD calling and TAD border strength assessment.
  - DI, IS, and CI are easy to compute: each of them can be calculated in $O(NK)$ time for one chromosome, where N is a number of bins in a chromosome, and $2K$ is a number of bins in the $2L$-vicinity of each bin. Typically, K is much less than N.
  - We need an arbitrary threshold / percentile or a kind of HMM to call TADs with these metrics.
  - There are almost no published and well-tested tools for TAD calling using these metrics.

Conclusion

- **TADs** are stable and evolutionary conserved units of transcription regulation in mammals. Some similar self-interacting domains were found in other Eukaryotic species.
- **Pros** and **cons** of **considered TAD calling methods**:
  - DI, IS, and CI are intuitive and inferred directly from TAD definition.
  - They can be used both for TAD calling and TAD border strength assessment.
  - DI, IS, and CI are easy to compute: each of them can be calculated in $O(NK)$ time for one chromosome, where N is a number of bins in a chromosome, and $2K$ is a number of bins in the $2L$-vicinity of each bin. Typically, K is much less than N.
  - We need an arbitrary threshold / percentile or a kind of HMM to call TADs with these metrics.
  - There are almost no published and well-tested tools for TAD calling using these metrics.
  - DI, IS, and CI can't enable us to call a TAD hierarchy (a TAD with its sub-TADs) as a whole.

Conclusion

- **Pros** and **cons** of **considered methods**:
  - DI, IS, and CI are intuitive and inferred directly from TAD definition.
  - They can be used both for TAD calling and TAD border strength assessment.
  - DI, IS, and CI are easy to compute: each of them can be calculated in $O(NK)$ time for one chromosome, where N is a number of bins in a chromosome, and $2K$ is a number of bins in the $2L$-vicinity of each bin. Typically, K is much less than N.
  - We need an arbitrary threshold / percentile or a kind of HMM to call TADs with these metrics.
  - There are almost no published and well-tested tools for TAD calling using these metrics.
  - DI, IS, and CI can't enable us to call a TAD hierarchy (a TAD with its sub-TADs) as a whole.
- **In Part 2** I'll consider *some* of the following much more complicated methods and tools for TAD calling: Sexton et al., 2012; Hou et al., 2012; Armatus, 2014; HiCseg, 2014; Arrowhead algorithm, 2014; TADtree, 2015; TADbit.

Outline

Chromatin conformation overviews

- Nguyen H. G. and Bosco G. 2015. Gene positioning effects on expression in Eukaryotes. *Annual Review of Genetics* 49: 627–646.
- Gibcus J. H. and Dekker J. 2013. The hierarchy of the 3D genome. *Molecular Cell* 49(5): 773–782.
- Dekker J. and Heard E. 2015. Structural and functional diversity of topologically associating domains. *FEBS Letters* 589(20, Part A): 2877–2884.

Self-interacting chromatin domains in various species

- **Chromatin interaction domains (CIDs) in bacterium *Caulobacter crescentus*:** Le T. B. et al. 2013. High-resolution mapping of the spatial organization of a bacterial chromosome *Science* 342(6159): 731–734.
- **Chromatin globules in *S. pombe*** Mizuguchi T. et al. 2014. Cohesin-dependent globules and heterochromatin shape 3D genome architecture in S. pombe . *Nature* 516(7531): 432–435.
- **Physical domains in *Drosophila:*** Sexton T. et al. 2012. Three-dimensional folding and functional organization principles of the Drosophila Genome. *Cell* 148(3): 458–472.
- **TADs in *C. elegans*** Crane E. et al. 2015. Condensin-driven remodeling of X-chromosome topology during dosage compensation. *Nature* 523(7559): 240–244.
- **TADs in human and mouse:** Dixon J. R. et al. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485(7398): 376–380.

Chromatin conformation capture methods:

- **Overview:** de Wit E. and de Laat W. 2012. A decade of 3C technologies: insights into nuclear organization. *Genes & Development* 26(1): 11–24.
- **Hi-C:** Lieberman-Aiden E. et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326(5950): 289–293.
- **Some Hi-C derivatives:**
  In-situ Hi-C: Rao S. S. et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159(7): 1665–1680.
  Capture Hi-C: Mifsud B. et al. 2015. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics* 47(6): 598–606.

Hi-C data processing and analysis

Overviews:

- Lajoie B. R. et al. 2015. The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods* 72: 65 – 75.
- Ay F. and Noble W. S. 2015. Analysis methods for studying the 3D architecture of the genome. *Genome Biology* 16:183.

Hi-C data correction:

- Imakaev M. et al. 2012. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods* 9(10): 999–1003.
- Yaffe E. and Tanay A. 2011. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genetics* 43(11): 1059–1065.

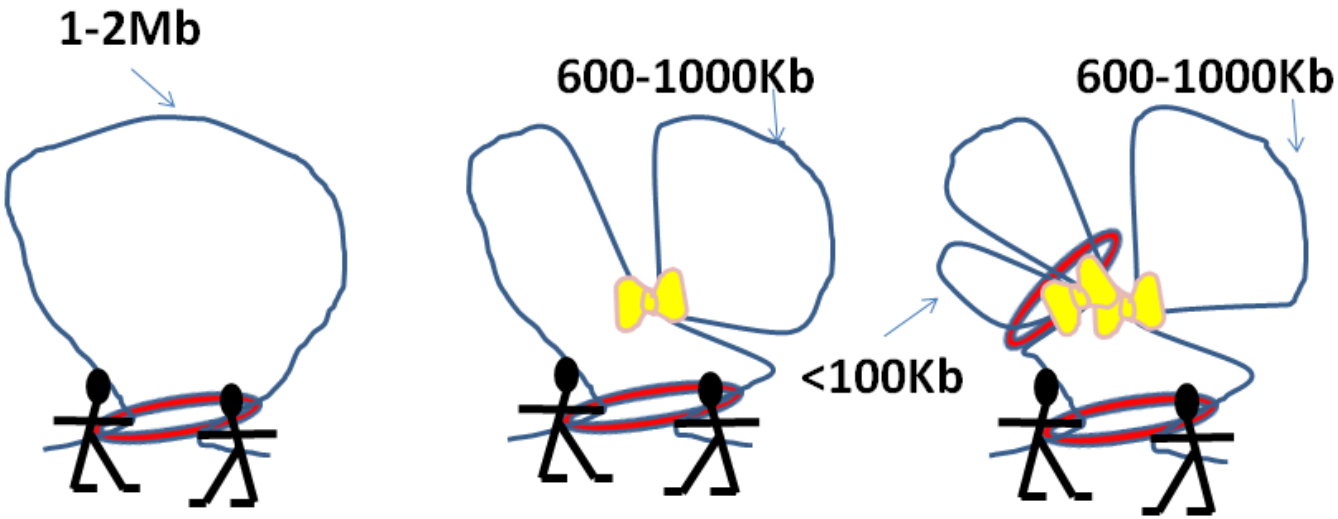
TAD calling methods

Covered in this overview:

- **Directionality index:** Dixon J. R. et al. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485(7398): 376–380.
- **Insulation score:** Crane E. et al. 2015. Condensin-driven remodeling of X-chromosome topology during dosage compensation. *Nature* 523(7559): 240–244.
- **Contrast index:** Van Bortle K. et al. 2014. Insulator function and topological domain border strength scale with architectural protein occupancy. *Genome Biology* 15(6): R82.
  Alekseyenko A. A. et al. 2015. The oncogenic BRD4-NUT chromatin regulator drives aberrant transcription within large topological domains *Genes & Development* 29(14): 1507–1523.

Additional: $\log_2$**-ratio:** Mizuguchi T. et al. 2014. Cohesin-dependent globules and heterochromatin shape 3D genome architecture in S. pombe *Nature* 516(7531): 432–435.

Thank you!



Sam Rose. Epigenetics and organisation