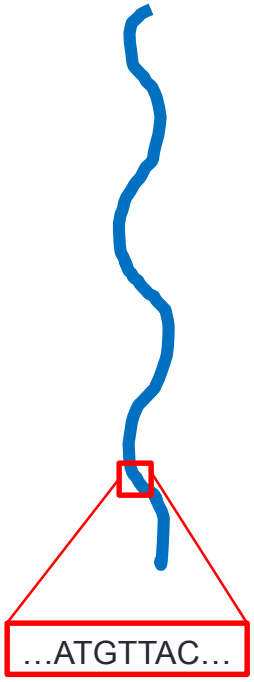
MULTIPLE LOGISTIC REGRESSION FOR TOPOLOGICAL CHROMATIN DOMAIN BORDER ANALYSIS

Raphaël MOURAD, Assist. Prof.,
Chromatin Dynamics and Cell Proliferation lab
University Paul Sabatier, Toulouse III

INTRODUCTION

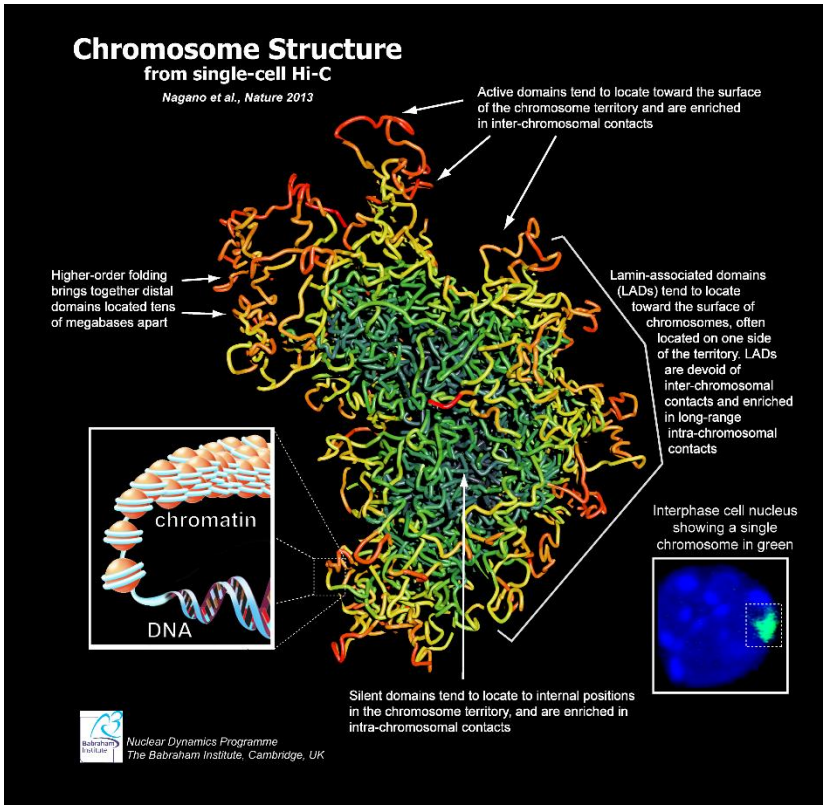
3D structure of chromosome

Chromosome
~ 10 cm long



DNA

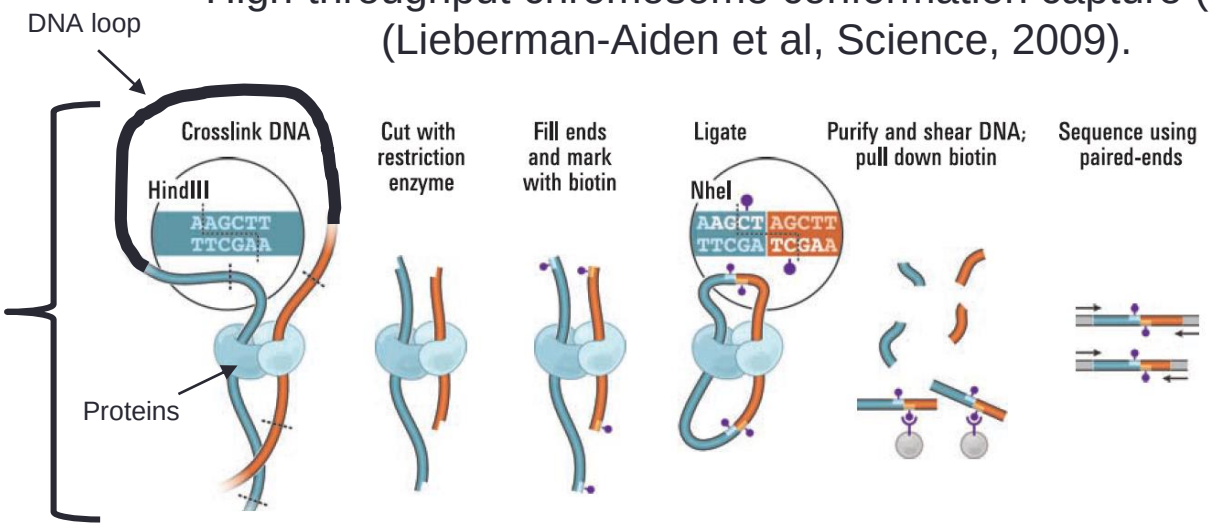
Compaction into
the cell nucleus
(5 µm)



3D chromosome structure assessment

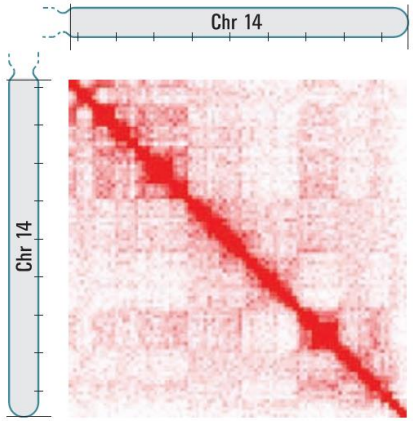
Next-Gen Sequencing Experiment

High-throughput chromosome conformation capture (Hi-C) (Lieberman-Aiden et al, Science, 2009).

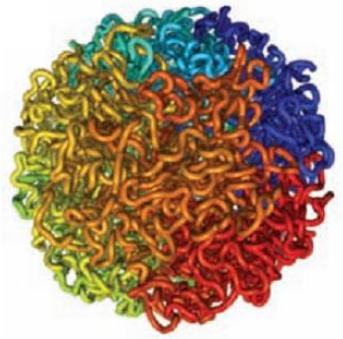


Results

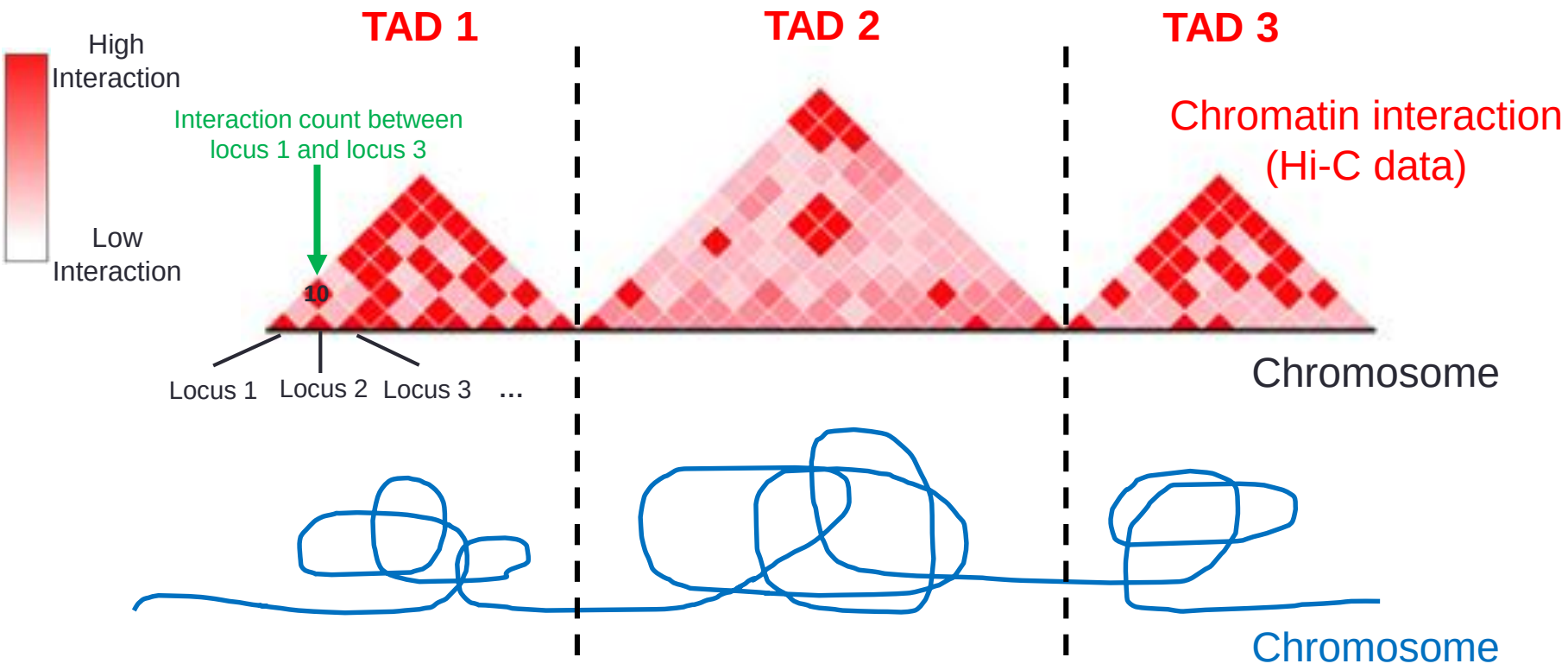
Interaction heatmap



3D structure of chromosome



Chromosomes are spatially structured in 3D domains

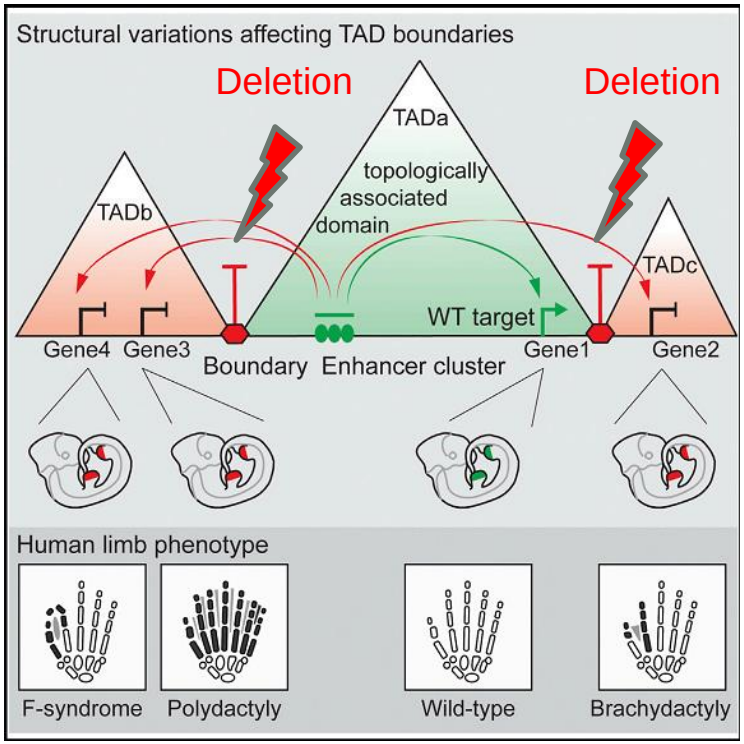


Chromosomes are spatially structured in topologically associating domains (TADs) (Sexton et al., Cell, 2012; Dixon et al., Nature, 2012).

TADs are stable across different cell types and highly conserved across species.

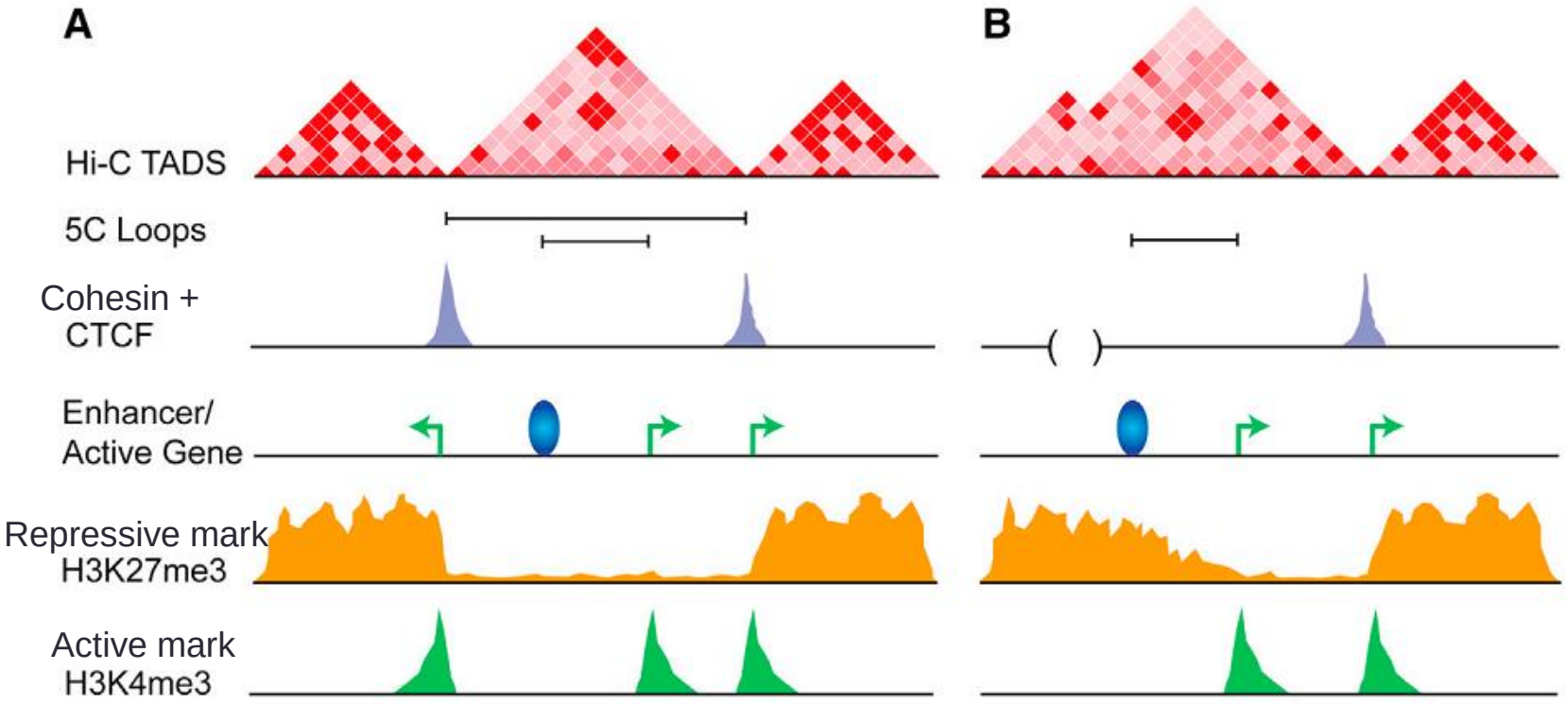
The biological role of these 3D domains

- 3D domains play an important role in:
  - gene expression regulation,
  - DNA replication
  - …
- For instance, loops between enhancers and promoters that regulate gene expression are constrained by 3D domains.
- Removal of these contraints by deletions of domain boundaries can cause de novo enhancer-promoter interactions and misexpression, and can lead to **genetic diseases**.



Lupianez et al., Cell, 2015.

Architectural proteins: Key drivers of 3D structure?

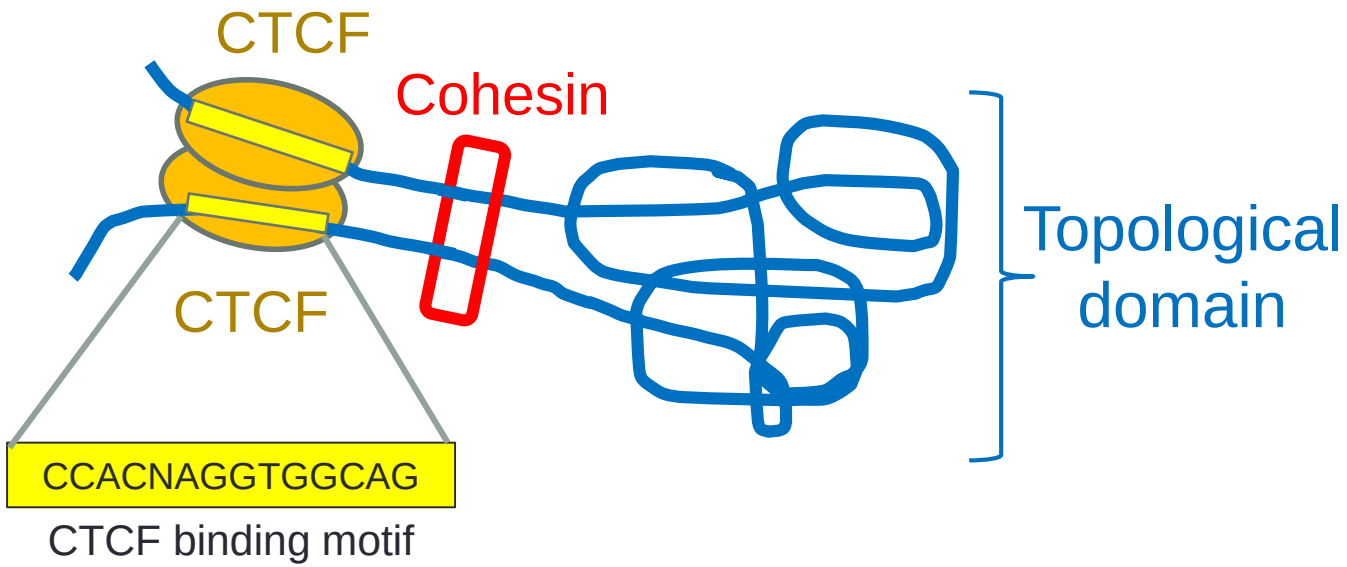


A. Model for CTCF + cohesin in demarcating TAD-borders in mammals.

B. Blurring of TAD boundaries after deletion of a border.

(Phillips-Cremins and Corces. Molecular Cell, 50(4):461-474, May 2013)

Architectural protein model in mammals



- In mammals, CTCF is thought to be the key insulator binding protein that works with cofactor cohesin to maintain 3D domain borders (Rao et al., Cell, 2015; Sanborn et al., PNAS, 2015).

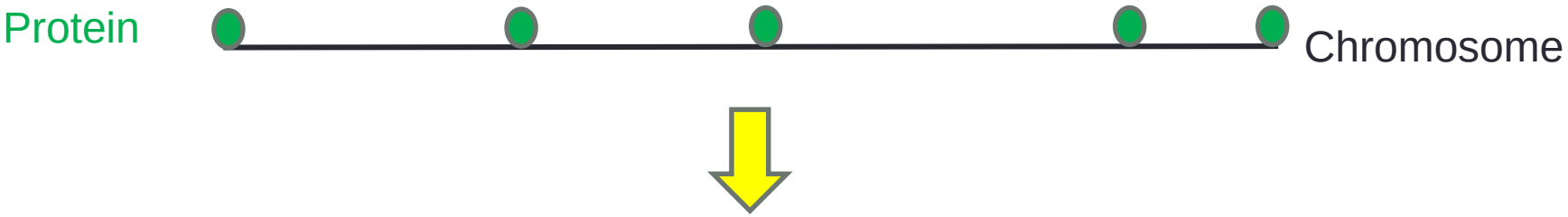
Architectural protein model in *Drosophila*

- In Drosophila, many insulator binding proteins (IBPs) colocalize with TAD borders:
  - dCTCF,
  - BEAF-32,
  - GAF,
  - Su(Hw),
  - dTFIIIC.
- And several cofactors are recruited by IBPs:
  - CP190, cohesin, chromator, condensin I/II and Fs(1)h-L.

Van Bortle et al., Genome Biology, 2015.

Li et al., Mol Cel, 2014.

How to identify architectural proteins? Univariate enrichment test



- Contingency table:

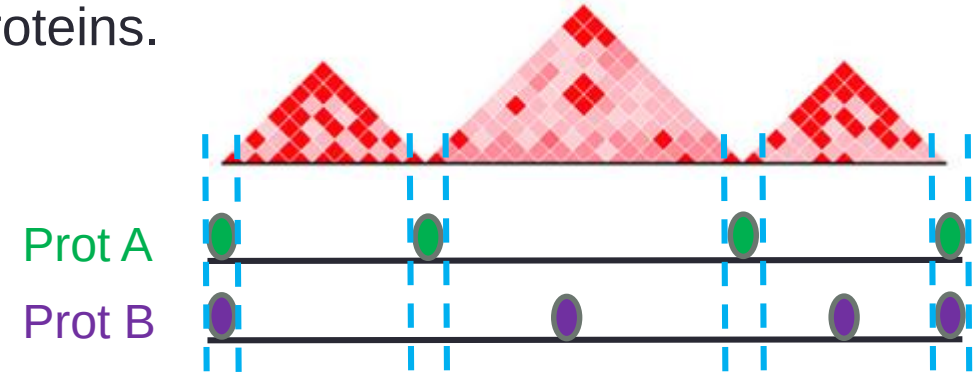
	Presence of the protein	Absence of the protein	Odds
Inside border	3	1	3/1=3
Outside border	2	24	2/24=0,08

How to identify architectural proteins? Univariate enrichment test

- From the previous contingency table, on can estimate the odds ratio (noted OR):
  - $\widehat{OR} = \frac{3/1}{2/24} = \frac{3}{0{,}0833} = 36$
- One can apply a Fisher's exact test to assess enrichment. There are two hypotheses about the odds ratio:
  - $H_0 : OR = 1$
  - $H_1 : OR \neq 1$
- OR reflects either enrichment ($OR > 1$) or impoverishment ($OR < 1$) of the architectural protein at TAD borders.

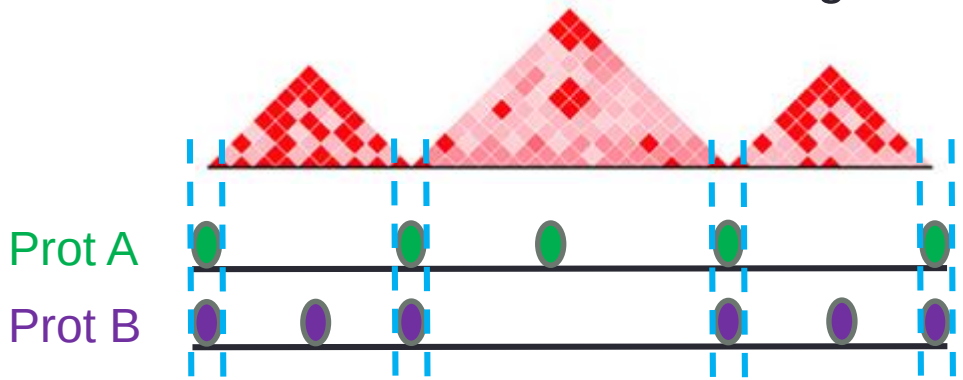
Caveats of univariate enrichment test

- Univariate enrichment test does not account for :
  - Potential correlations (*i.e.*, colocalizations) among the architectural proteins.



Both proteins A and B colocalize to borders, but only protein A truly influences borders.

  - Potential statistical interactions among the architectural proteins.



Proteins A and B need to be together to influence borders.

PROPOSED APPROACH:
MULTIPLE LOGISTIC REGRESSION

Logistic regression formulation of univariate enrichment test

- The previous univariate enrichment test can be reformulated as a logistic regression model:
  - $ln\frac{Prob(Y=1|X)}{1-Prob(Y=1|X)} = \beta_0 + \beta X$
  - Variable Y indicates if the genomic bin belongs to the boundary (Y = 1) or if the genomic bin is outside of the boundary (Y = 0).
  - Variable X can :
    - either denotes the presence (X = 1) or the absence (X = 0) of the protein within the genomic bin,
    - or quantify ChIP-seq signal intensity within the genomic bin ($\log\left({ChIP}/{Input}\right)$).

Logistic regression formulation of univariate enrichment test

- In the logistic regression, the slope parameter β is the natural logarithm of the abovementioned odds ratio OR.

- Parameter β of the logistic regression model reflects enrichment ($\beta > 0$) or depletion ($\beta < 0$) of the architectural protein at TAD borders.

- Parameter β can be tested by a Wald test:
  - $W = \frac{\hat{\beta}}{\hat{\sigma}_{\beta}}$
  - $\hat{\sigma}_{\beta}$ denotes the standard error of parameter $\hat{\beta}$.
  - Statistic W follows a normal distribution.
  - In practice, we observed that Wald test yields p-values that are similar to the ones obtained from the often preferred Likelihood Ratio Test.

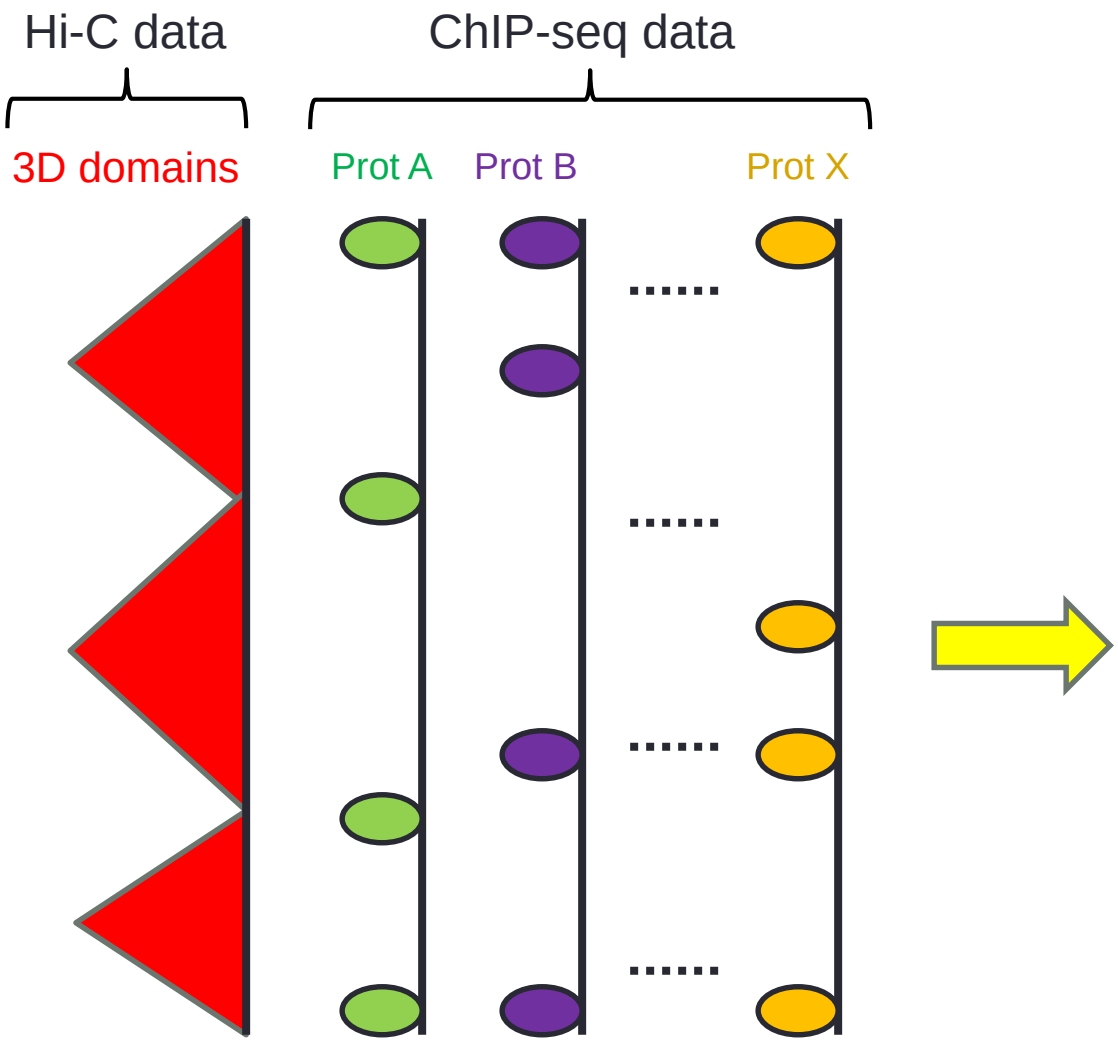
Multiple logistic regression

- Logistic regression model provides a natural framework for analysis over p genomic features:
  - $ln\frac{Prob(Y=1|X)}{1-Prob(Y=1|X)} = \beta_0 + \boldsymbol{\beta X}$
  - Where $\boldsymbol{X} = \{X_1, \ldots, X_p\}$ is the set of p proteins of interest and $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_p\}$ denotes the set of corresponding slope parameters (one parameter β for each protein).

- Here we assess in the same model all the architectural proteins of interest!

- We thus account for potential colocalizations among the proteins (i.e. conditional independence).

Multiple logistic regression data



Model variables

Y	X1	X2	…	Xp
1	1	1	…	1
0	0	0	…	0
0	0	1	…	0
0	0	0	…	0
1	1	0	…	0
0	0	0	…	0
0	0	0	…	1
0	0	0	…	0
0	0	1	…	1
1	1	0	…	0
0	0	0	…	0
0	0	0	…	0
1	1	1	…	1

Statistical interaction

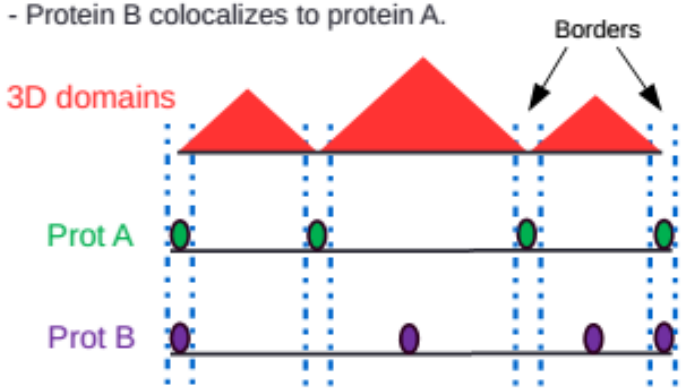
- Interaction terms can be included in the logistic regression model to account for potential interactions between genomic features.

- For instance, one can include in the model an interaction term between two proteins X_1 and X_2:
  - $ln \frac{Prob(Y=1|X)}{1-Prob(Y=1|X)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2$
  - The product $X_1 X_2$ is the statistical interaction term between the two proteins X_1 and X_2.
  - Parameter β_{12} measures the enrichment of interaction $X_1 X_2$.

Illustration of the model

Scenario 1 (no interaction):

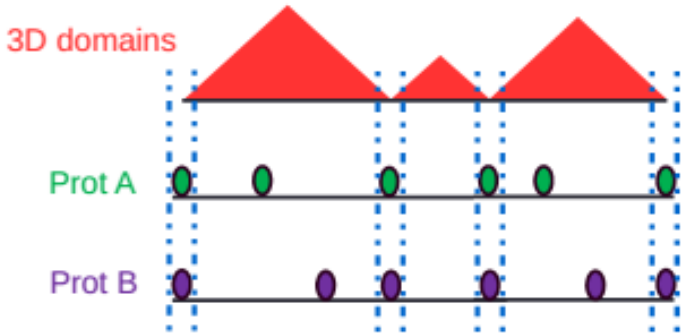
- Protein A influences 3D domain borders.
- Protein B colocalizes to protein A.



Enrichment test	Multiple logistic regression
$\beta_A > 0$ Prot A is enriched.	$\beta_A > 0$ **Prot A influences borders.**
$\beta_B > 0$ Prot B is enriched.	$\beta_B = 0$ Prot B does not influence borders.

Scenario 2 (interaction):

- The co-occurrence of proteins A and B influences 3D domain borders, but not the proteins alone.



Enrichment test	Multiple logistic regression
$\beta_A > 0$ Prot A is enriched.	$\beta_A = 0$ Prot A does not influence borders.
$\beta_B > 0$ Prot B is enriched.	$\beta_B = 0$ Prot B does not influence borders.
$\beta_{AB} > 0$ Interaction between prot A and B is enriched.	$\beta_{AB} > 0$ **Interaction between prot A and B influences borders.**

RESULTS

Analysis of architectural proteins in *Drosophila*

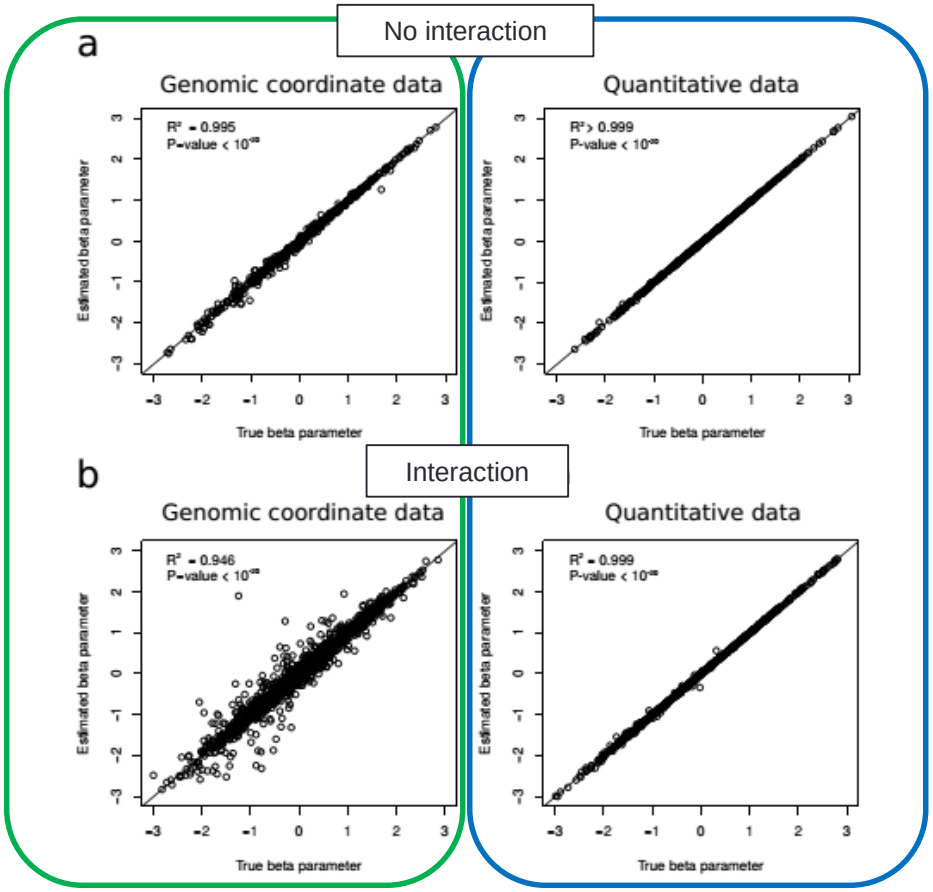
- We illustrate the multiple logistic regression with *Drosophila melanogaster*.

- Fly represents an interesting model to study architectural proteins because there are many insulator binding proteins.

- The data:
  - ChIP-seq data from Kc167 cells (Corces *et al*.),
  - Hi-C data from Kc167 cells (Corces *et al*.).

Parameter estimation accuracy

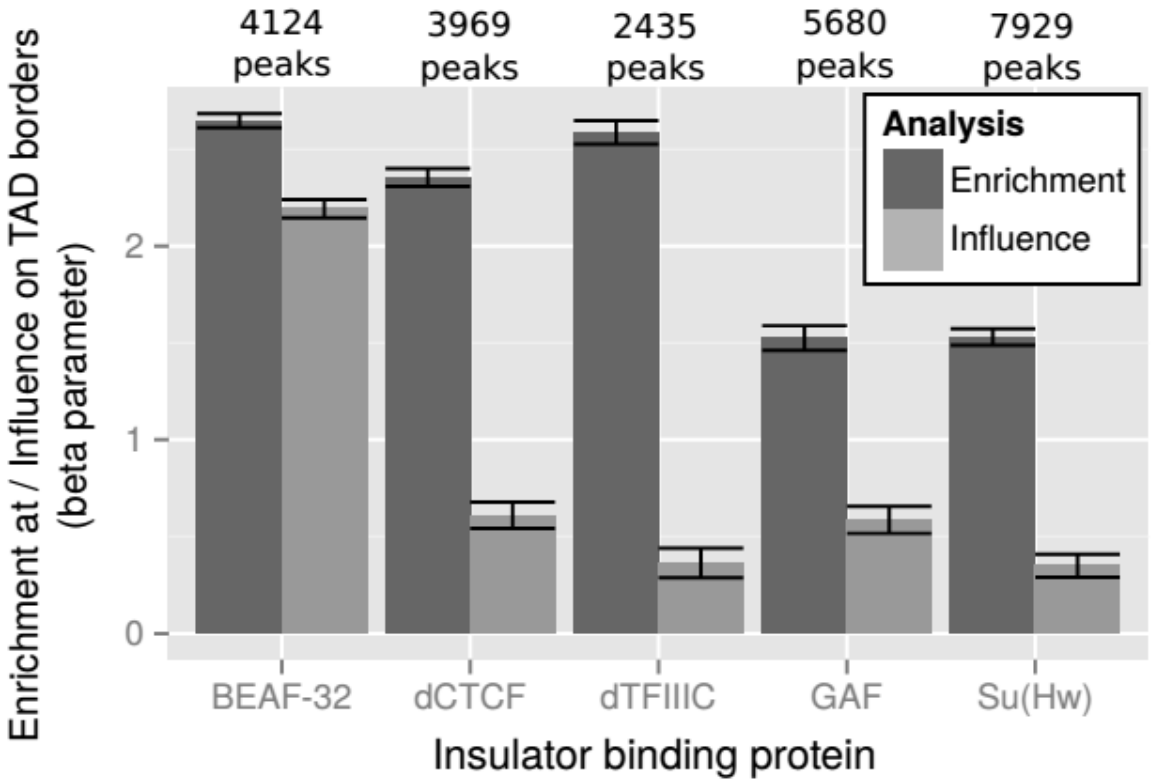
Qualitative X:
Presence
or absence of
the protein



Quantitative X:
ChIP-seq signal
$\log\left({ChIP}/{Input}\right)$

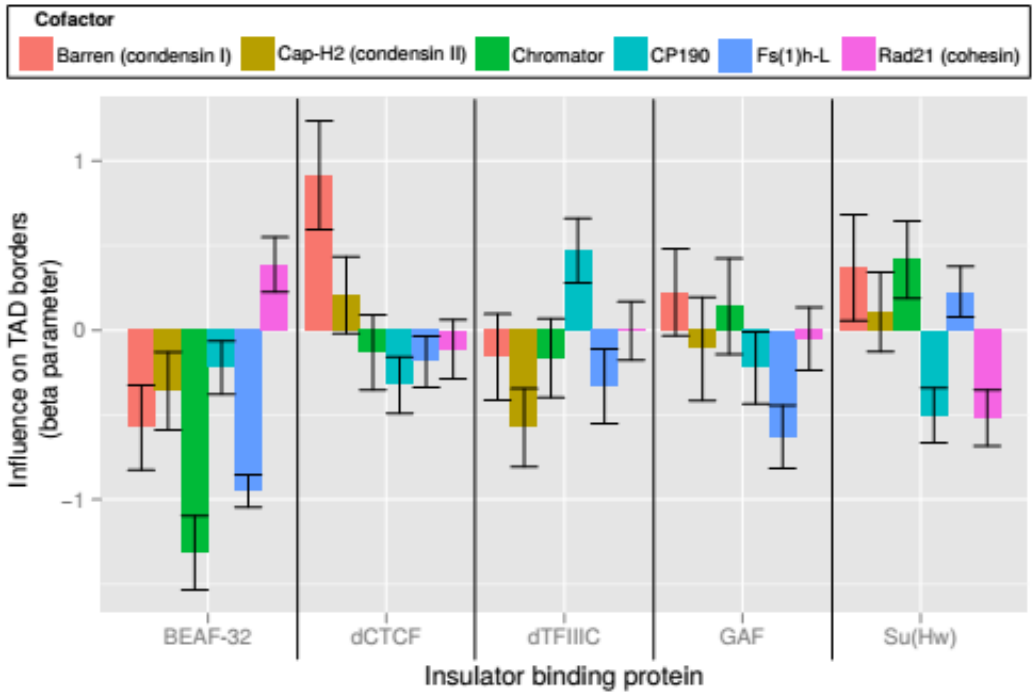
Very good parameter estimation accuracies are achieved for both marginal variables with $R^2 = 99{,}6\%$ (no interactions) and two-way interaction variables with $R^2 = 94{,}6\%$.

Analysis of insulator binding proteins



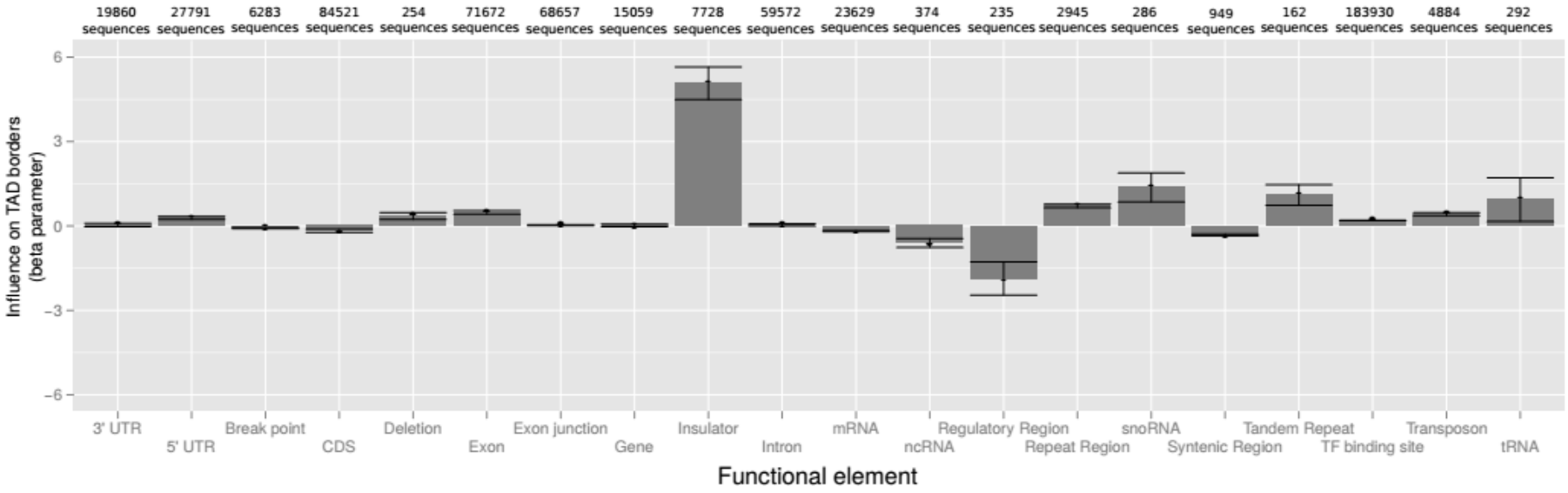
Although all these insulator binding proteins (IBPs) are enriched at TAD borders, only BEAF-32 highly impacts TAD demarcation establishment or maintenance.

Analysis of statistical interactions



- Significant positive interactions reflected synergistic effects of IBPs with cofactors, especially for dCTCF with condensin I.
- Significant negative interactions revealed antagonistic effects at domain borders, in particular for BEAF-32 with cofactors Chromator and with Fs(1)h-L.

Analysis of functional elements



- Insulators were by far the most influential functional elements with respect to domain borders, as established in human (Rao et al., Cell 2014).
- We found positive effects for repeat regions and for snoRNA genes.
- A negative impact on TAD border was detected for regulatory sequences.

Analysis of DNA binding proteins in human

- We illustrate the multiple logistic regression with GM12878 cells.
- For GM12878 cells, there are very high resolution Hi-C data at 1kb that allowed to accurately identify 3D domains.
- The data:
  - ChIP-seq data from GM12878 cells (ENCODE project),
  - Hi-C data from GM12878 cells (Rao et al., Cell, 2014).

Results

- CTCF and cohesin (subunit Rad21) presented the highest effects among all factors (CTCF: $\hat{\beta} = 1{,}90$; cohesin: $\hat{\beta} = 1{,}91$), in complete agreement with numerous studies (Rao et al., Cell 2014).

- Our model also detected large positive effects for ZNF143 ($\hat{\beta} = 1{,}85$) and for EZH2, the catalytic subunit of the Polycomb repressive complex 2, ($\hat{\beta} = 1{,}32$) in total agreement with a very recent studies (Bailey et al., Nat. Comm. 2015; Schoenfelder et al., Nat. Genet. 2015).

- In addition, our model revealed several factors associated with transcriptional activation that had significant negative influences on TAD borders. These proteins included RXRA ($\hat{\beta} = -1{,}37$), P300 ($\hat{\beta} = -1{,}22$), BCL11A ($\hat{\beta} = -0{,}82$) and ELK1 ($\hat{\beta} = -0{,}74$).

CONCLUSION AND PERSPECTIVES

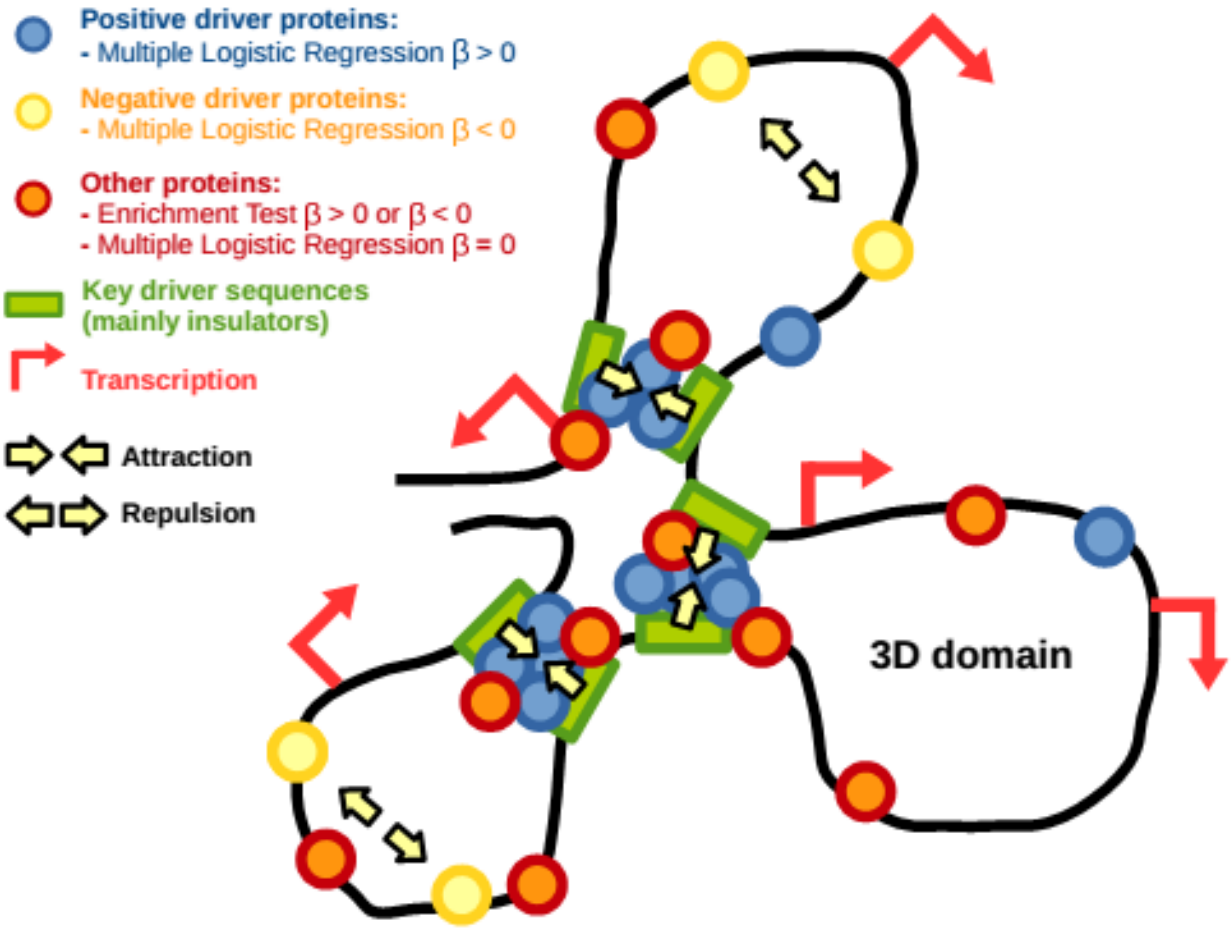
Conclusion

- Here, we describe a multiple logistic regression (MLR) to assess the roles of genomic features such as DNA binding proteins and functional elements on TAD border establishment or maintenance.

- Using simulations, we show that model parameters can be accurately estimated for both marginal genomic features (no interaction) and two-way interactions.

- Using experimental *Drosophila* Hi-C and ChIP-seq data, we show that the proposed model can identify genomic features that are most influential with respect to TAD borders.

A new model for 3D domain border establishment or maintenance



Future directions

- In this article, we have focused on the influences of proteins on 3D domain border establishment or maintenance.
- Another important question is to understand the role of DNA-binding proteins in chromatin interactions within domains.
- For instance, it is essential to identify proteins that influence interactions between enhancers and promoters that regulate gene expression.

Bibliography

- Erez Lieberman-Aiden, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science, 326(5950):289-293, October 2009.
- Tom Sexton, et al. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell, 148(3):458-472, February 2012.
- Chunhui Hou, et al. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. Molecular Cell, 48:471-484, November 2012.
- Jesse R. Dixon, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature, 485(7398):376-380, May 2012.
- Jennifer E. Phillips-Cremins and Victor G. Corces. Chromatin insulators: Linking genome organization to cellular function. Molecular Cell, 50(4):461-474, May 2013.
- Jun Liang, et al. Chromatin immunoprecipitation indirect peaks highlight functional long-range interactions among insulator proteins and RNAII pausing. Molecular Cell, 53(4):672-681, February 2014.
- Kevin Van Bortle, et al. Insulator function and topological domain border strength scale with architectural protein occupancy. Genome Biology, 15(5):R82+, June 2014.
- Li Li, et al. Widespread rearrangement of 3D chromatin organization underlies Polycomb-mediated stress-induced silencing. Molecular Cell, (15):S1097-2765, March 2015.
- Suhas S. P. Rao, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell, 159(7):1665-1680, February 2015.
- D. G. Lupianez et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell, 161(5):1012-1025, May 2015.