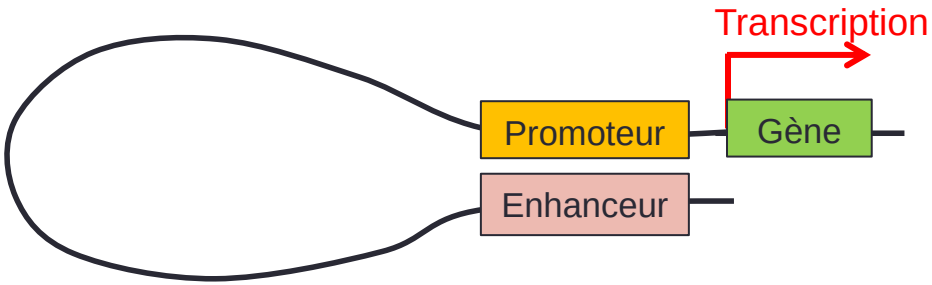
ETUDE DU LIEN ENTRE PROTEINES ET BOUCLES ADN PAR REGRESSION AVEC TERMES D'INTERACTION

Raphaël MOURAD, Maître de conférences
Laboratoire de Biologie Moléculaire Eucaryote
Université Paul Sabatier, Toulouse III

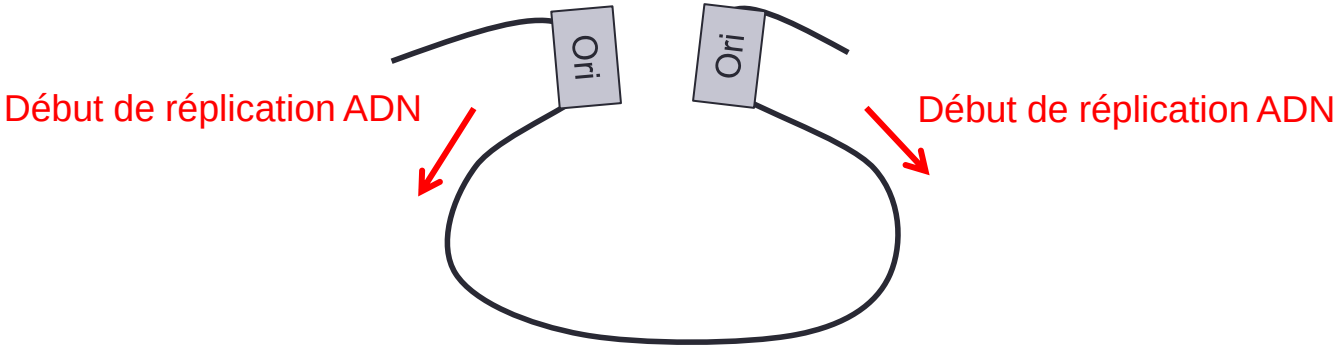
INTRODUCTION

Rôle de l'organisation 3D du génome dans la cellule

- Régulation de l'expression des gènes via les contacts de longue distance entre enhanceurs et promoteurs de gènes cibles.

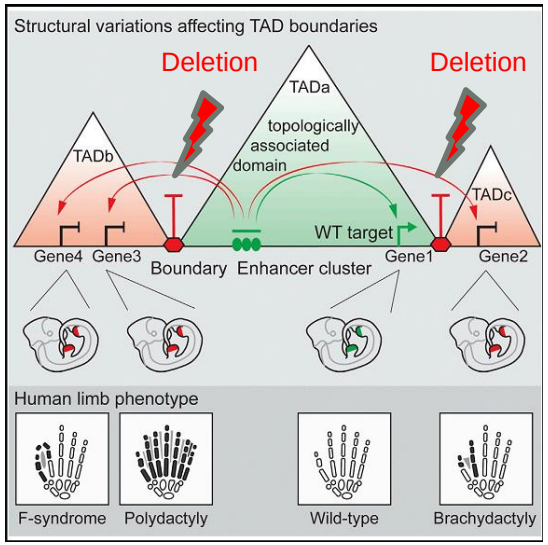


- Régulation du timing de la réplication de l'ADN.



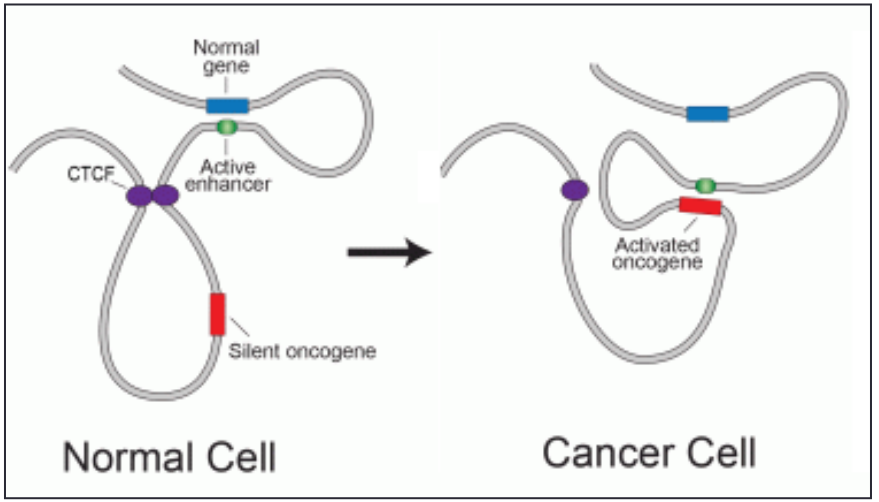
Impact de l'organisation 3D sur l'apparition de maladies

- La délétion de frontières entre les domaines 3D peut entrainer de nouvelles interactions enhanceur-promoteur, une mauvaise expression de gènes cibles, et causer ainsi des **maladies génétiques**.



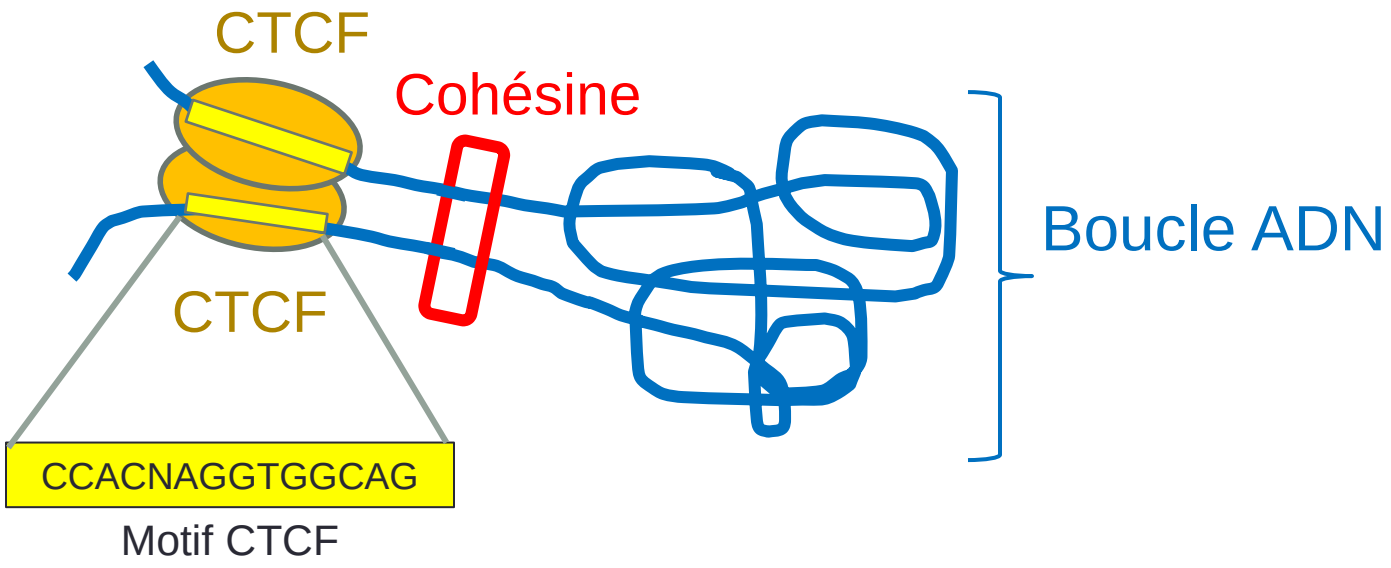
Lupianez et al., Cell 2015.

- La délétion de frontières entre les domaines 3D peut entrainer l'activation de proto-oncogènes en oncogènes, responsables de nombreux **cancers**.



Hnisz et al., Science 2016.

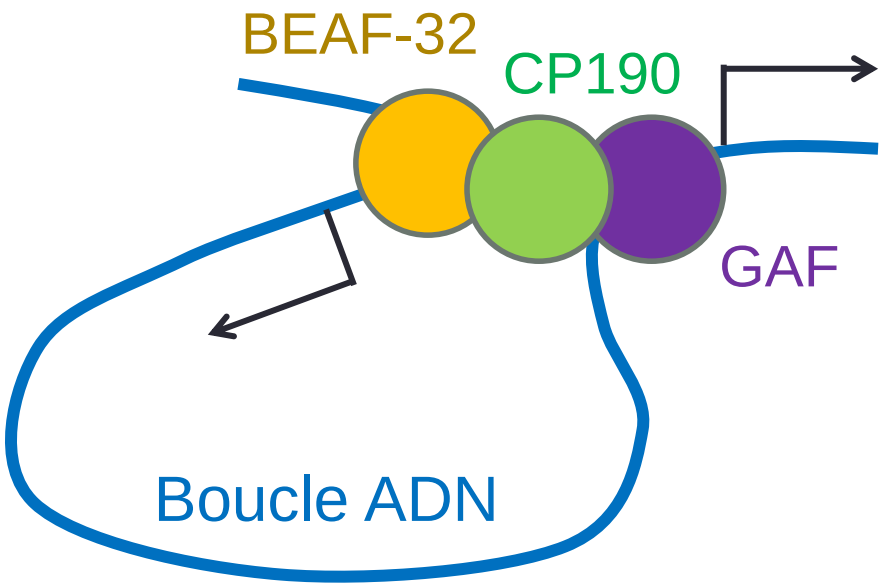
Mécanismes et facteurs moléculaires impliqués chez les mammifères



Rao et al., Cell 2015; Sanborn et al., PNAS 2015.

- CTCF: protéine insulatrice (= se fixant sur les séquences insulatrices).
- Cohésine: protéine cofacteur (= recrutée par CTCF).

Mécanismes et facteurs moléculaires impliqués chez la drosophile



Liang et al., Mol Cell 2014 (notre équipe)

- BEAF-32/GAF: protéines insulatrices.
- CP190: protéine cofacteur (= recrutée par BEAF-32 et GAF).

Cependant seule une partie des mécanismes et facteurs est connue!

- Les mécanismes présentés **n'expliquent qu'une partie de l'organisation 3D des chromosomes** et de son lien avec l'expression du génome.

- Chez l'homme, **d'autres facteurs ont récemment été révélés**, notamment ZNF143, PcG, mais aussi les marques histones tels que H3K27me3, H3K4me3….

- Chez la drosophile, beaucoup d'autres facteurs ont probablement un rôle tels que Chromator, DREF ou les condensines I/II.

Quels types de facteurs pouvant être impliqués?

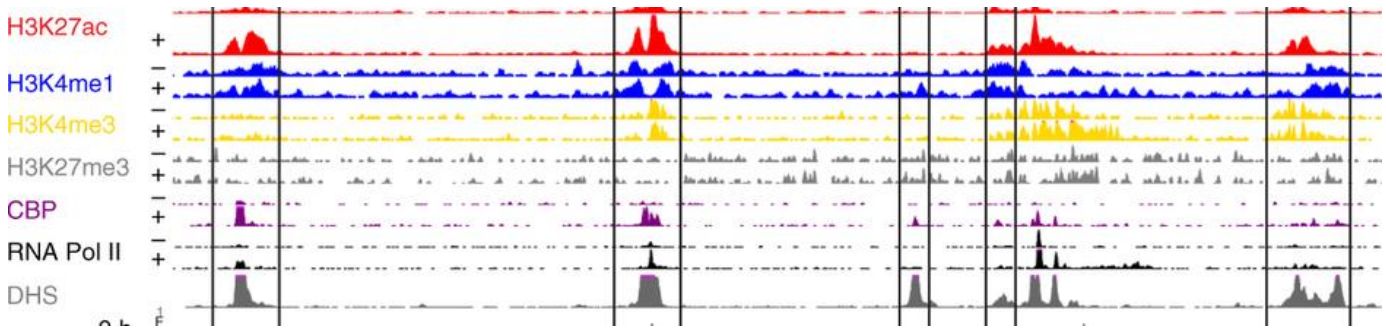
- Quels sont les protéines insulatrices, cofacteurs et motifs ADN impliqués dans l'organisation "**structurelle**" 3D du génome?

- Quels sont les facteurs de transcription impliqués dans les contacts "**fonctionnels**" enhanceur-promoteur? Quel est leur impact sur l'expression des gènes et/ou la pause de l'ARN Polymérase II?

Mais un grand nombre de données publiques disponibles pour répondre à la question!

- Des milliers de données de séquencage dans GEO, ENCODE et modENCODE.

- Il y a une accumulation des données Hi-C, ChIP-seq, DNase-seq, RNA-seq ainsi que les données d'annotations de gènes, promoteurs, enhanceurs, insulateurs…



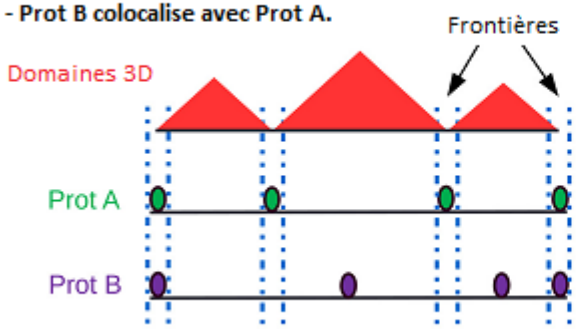
De nombreux challenges pour la bioinformatique et la biostatistique

- L'analyse des contacts nécessite l'utilisation de:
  - Modèles de régression pénalisée pour la sélection de variables;
  - Méthodes de machine learning pour la prédiction;
  - Réseaux d'interactions de type modèles graphiques probabilistes;
  - Analyse bioinformatique des motifs;

- Les données en jeu sont :
  - De grande taille (millions/milliards d'observations) et de grande dimension (des milliers de facteurs à tester);
  - Hétérogènes (intégration de données Hi-C, ChIP-seq, RNA-seq, annotations, motifs…).
  - Incertaines, bruitées et biaisées.

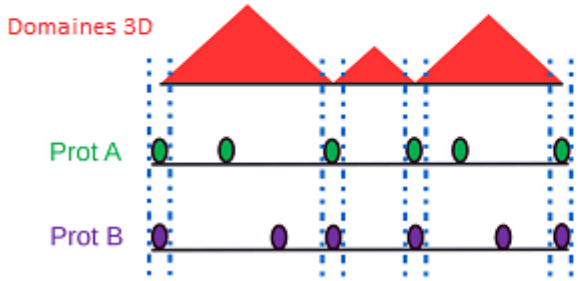
Travaux récents dans le domaine

Scenario 1 (pas d'interaction):
- Prot A influence les frontières des domaines 3D.
- Prot B colocalise avec Prot A.



Test d'enrichissement	Régression logistique multiple
$\beta_A>0$ Prot A enrichie.	$\beta_A>0$ **Prot A influence les frontières.**
$\beta_B>0$ Prot B enrichie.	$\beta_B=0$ Prot B n'influence pas les frontières.

Scenario 2 (interaction):
- la co-occurrence des prot A et B influence les frontières, mais pas les protéines seules.



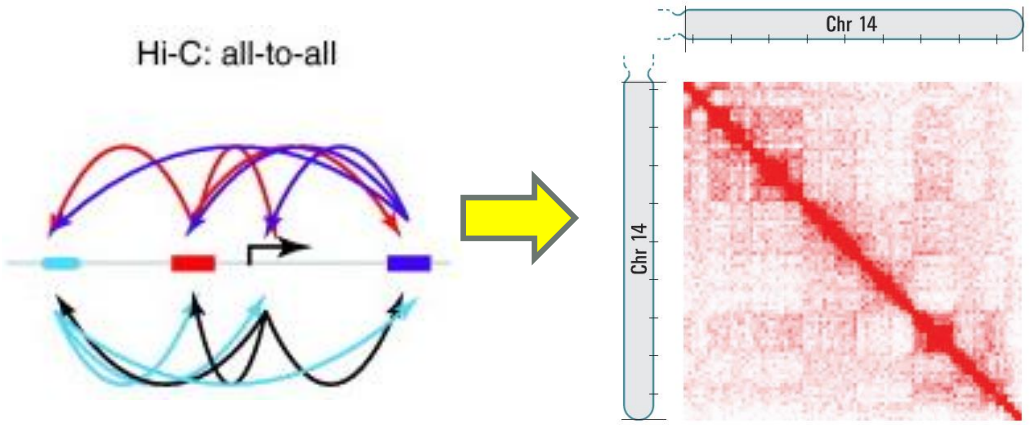
Test d'enrichissement	Régression logistique multiple
$\beta_A>0$ Prot A enrichie.	$\beta_A=0$ Prot A n'influence pas les frontières.
$\beta_B>0$ Prot B enrichie.	$\beta_B=0$ Prot B n'influence pas les frontières.
$\beta_{AB}>0$ Interaction entre prot A et B enrichie.	$\beta_{AB}>0$ **L'interaction entre prot A et B influence les frontières.**

Mourad et Cuvier, PLoS Comp Bio, 2016.

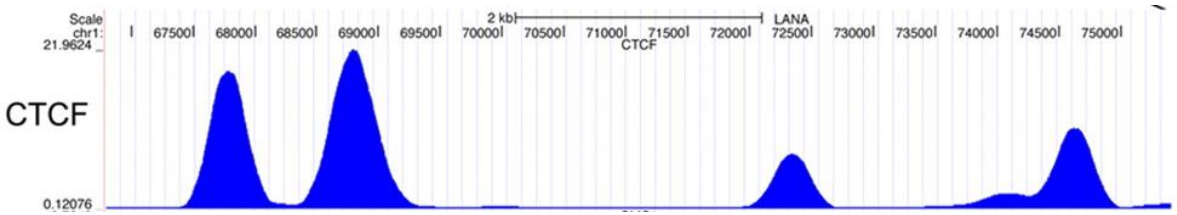
IDENTIFICATION DES FACTEURS IMPLIQUES DANS LES CONTACTS 3D

Données haut-débit et hétérogènes

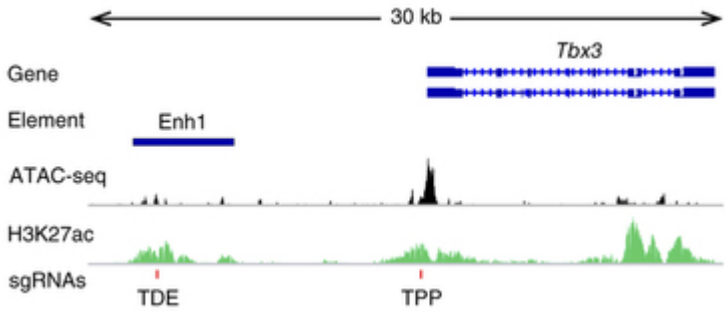
Données Hi-C
(et ChIA-PET, 5C,
Hi-C capture…)



Données ChIP-seq



Annotations
(enhanceurs, promoteurs,
insulateurs, motifs, …)



Un problème complexe

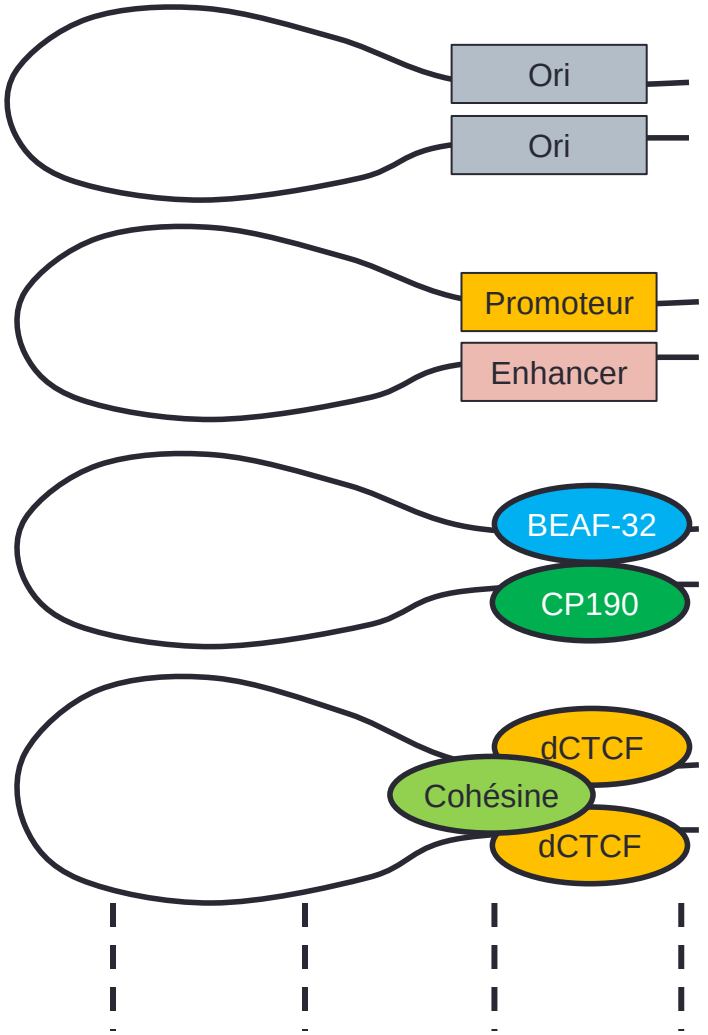
- Un grand nombre d'interactions statistiques entre facteurs à analyser.

- Un problème quadratique

(si interactions entre 2 facteurs).

- Un problème cubique

(si interactions entre 3 facteurs).



Approches existantes

- Test d'enrichissement de couples de facteurs dans les bins en contacts significatifs (Ka-Chun Wong et al., Bioinformatics 2015).
- Mesure de l'importance de variables (variable importance) dans la prédiction de contacts par Random Forests (He et al., PNAS 2014).
- Corrélation de la présence/absence des facteurs entre bins en contacts significatifs (Pancaldi et al., Genome Biology 2016; Zhang et al., Nature Comm 2016).

Le modèle proposé:
Modèle linéaire généralisé avec interactions

$$\log\left(\mathrm{E}\left[\mathbf{y}|\mathbf{X}\right]\right) = \beta_0 + \boldsymbol{\beta}\mathbf{X}$$
$$= \beta_0 + \beta_d\mathbf{d} + \boldsymbol{\beta}_B\mathbf{B} + \boldsymbol{\beta}_C\mathbf{C} + \beta_g\mathbf{g}$$

- Variables:
  - y : nombre de contacts entre deux bins,
  - d : log-distance entre bins (effet du polymère),
  - B={len,GC,map} : biais des données Hi-C,
  - C : facteurs de confusion (confounding variables),

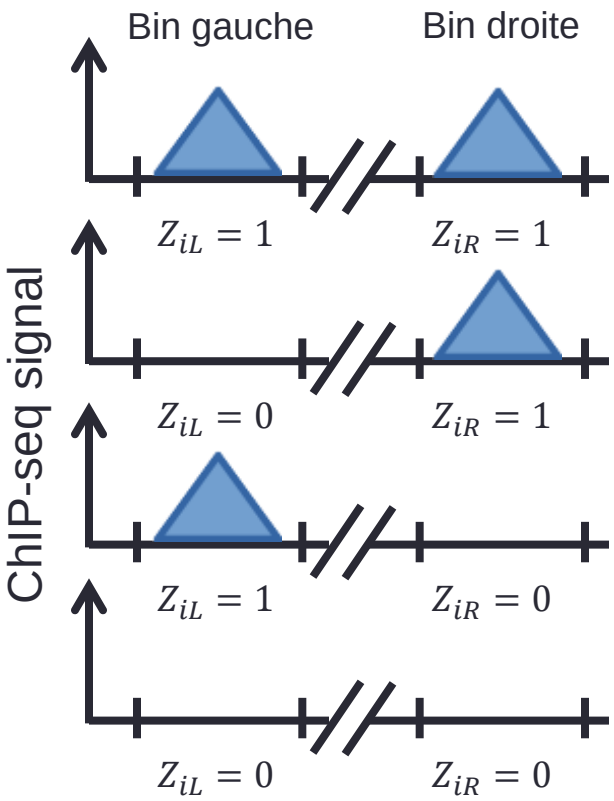
  - g : la variable génomique d'intérêt.

Variable $\mathrm{g} = n_{ii}$

$$\log\left(\mathrm{E}[\mathbf{y}|\mathbf{X}]\right) = \beta_0 + \beta_d\mathbf{d} + \boldsymbol{\beta}_B\mathbf{B} + \boldsymbol{\beta}_C\mathbf{C} + \beta_g\mathbf{g}$$
$$= \beta_0 + \beta_d\mathbf{d} + \boldsymbol{\beta}_B\mathbf{B} + \beta_{m_i}\mathbf{m}_i + \beta_{n_{ii}}\mathbf{n}_{ii}$$

Avec:

$$\mathbf{m}_i = \tfrac{1}{2}(\mathbf{z}_{iL} + \mathbf{z}_{iR})$$



Interaction:

$$n_{ii} = z_{iL} \times z_{iR}$$

Valeur de n_{ii}

$n_{ii} = 1$

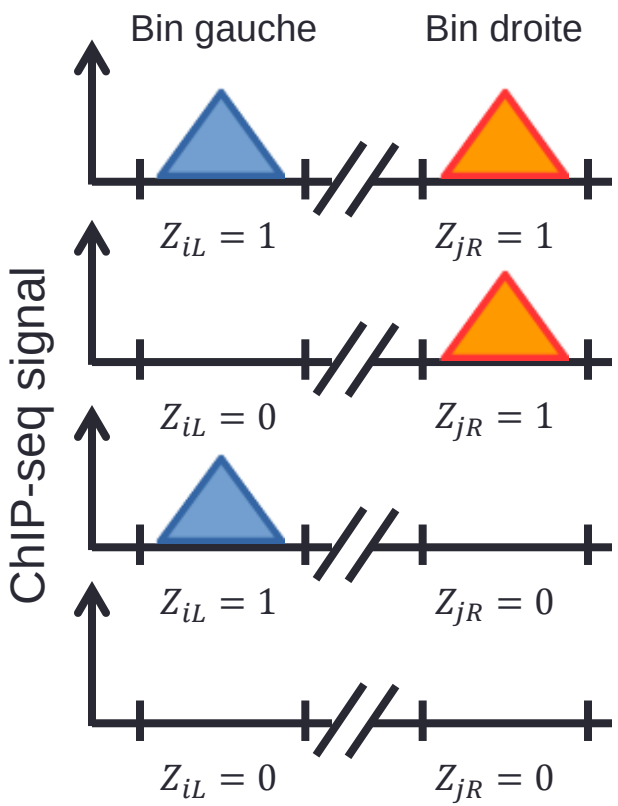
$n_{ii} = 0$

$n_{ii} = 0$

$n_{ii} = 0$

Variable $\mathrm{g} = n_{ij}$

$$\log\left(\mathrm{E}\left[\mathbf{y}|\mathbf{X}\right]\right) = \beta_0 + \beta_d\mathbf{d} + \boldsymbol{\beta}_B\mathbf{B} + \boldsymbol{\beta}_C\mathbf{C} + \beta_g\mathbf{g}$$
$$= \beta_0 + \beta_d\mathbf{d} + \boldsymbol{\beta}_B\mathbf{B} + \beta_{m_i}\mathbf{m}_i + \beta_{m_j}\mathbf{m}_j + \beta_{n_{ij}}\mathbf{n}_{ij}$$



Interaction:

$$n_{ij} = \frac{1}{2}\left(z_{iL} \times z_{jR} + z_{jL} \times z_{iR}\right)$$

Valeur de n_{ij}

$n_{ij} = 1$

$n_{ij} = 0$

$n_{ij} = 0$

$n_{ij} = 0$

Variables + complexes: g = c_{iik} et g = c_{ijk}

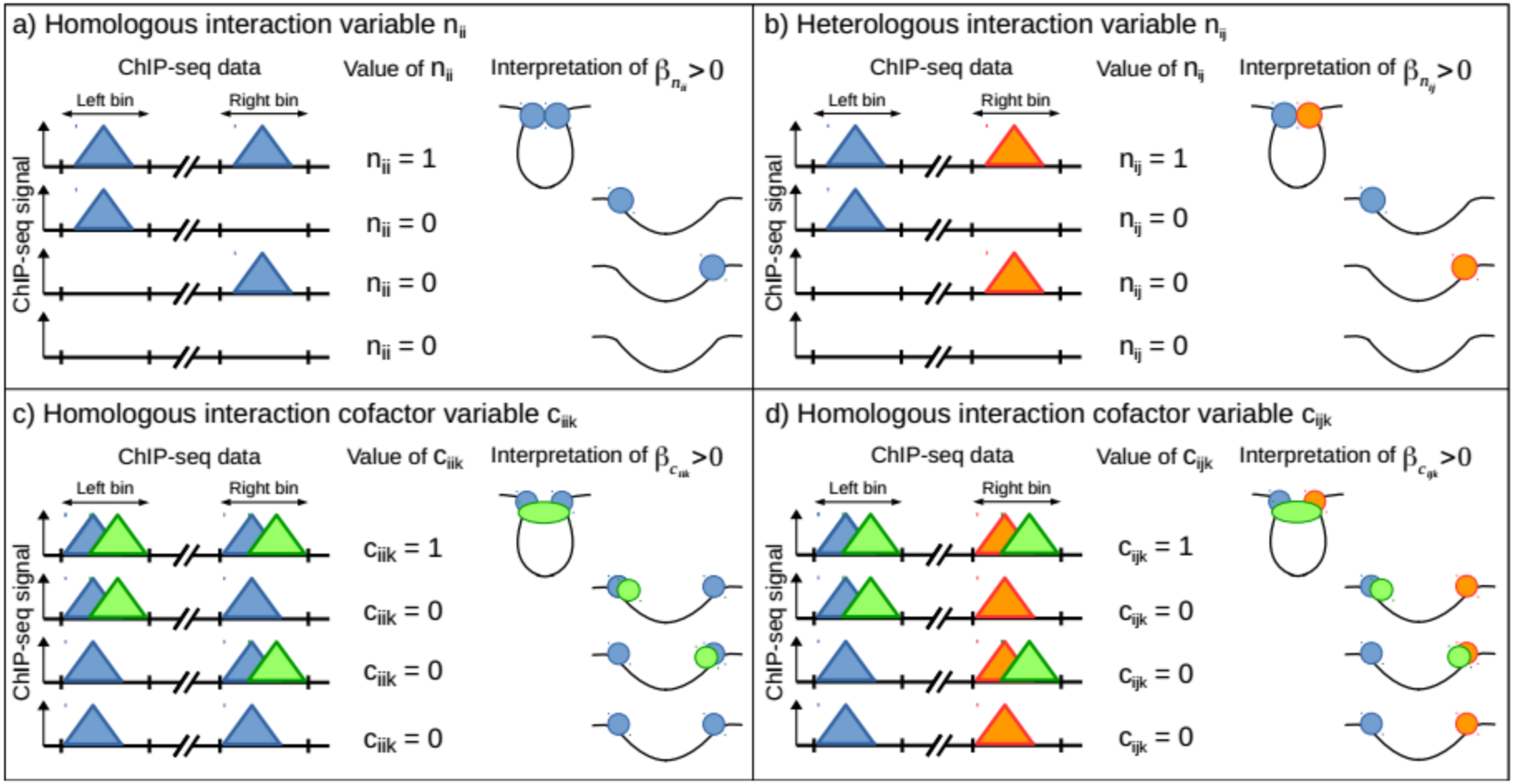
Modèle pour g = $c_{iik} = n_{ii} \times n_{kk}$

$$\begin{aligned}\log\left(\mathrm{E}[\mathbf{y}|\mathbf{X}]\right) &= \beta_0 + \beta_d\mathbf{d} + \boldsymbol{\beta}_B\mathbf{B} + \boldsymbol{\beta}_C\mathbf{C} + \beta_g\mathbf{g} \\ &= \beta_0 + \beta_d\mathbf{d} + \boldsymbol{\beta}_B\mathbf{B} + \beta_{m_i}\mathbf{m}_i + \beta_{m_k}\mathbf{m}_k + \beta_{m_{ik}}\mathbf{m}_{ik} + \beta_{n_{ii}}\mathbf{n}_{ii} + \beta_{n_{kk}}\mathbf{n}_{kk} + \beta_{n_{ik}}\mathbf{n}_{ik} \\ &+ \beta_{n_{ii}\times m_k}(\mathbf{n}_{ii} \times \mathbf{m}_k) + \beta_{n_{kk}\times m_i}(\mathbf{n}_{kk} \times \mathbf{m}_i) + \beta_{c_{iik}}\mathbf{c}_{iik},\end{aligned}$$

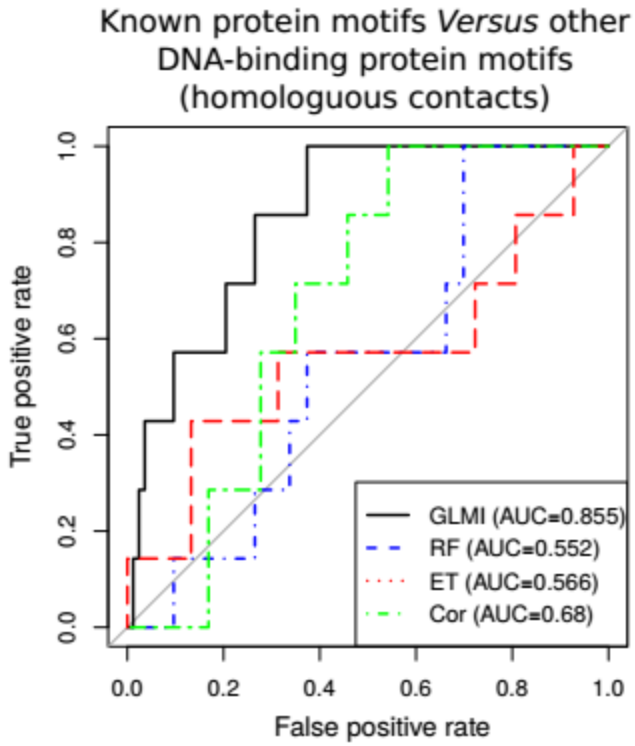
Modèle pour g = $c_{ijk} = n_{ij} \times n_{kk}$

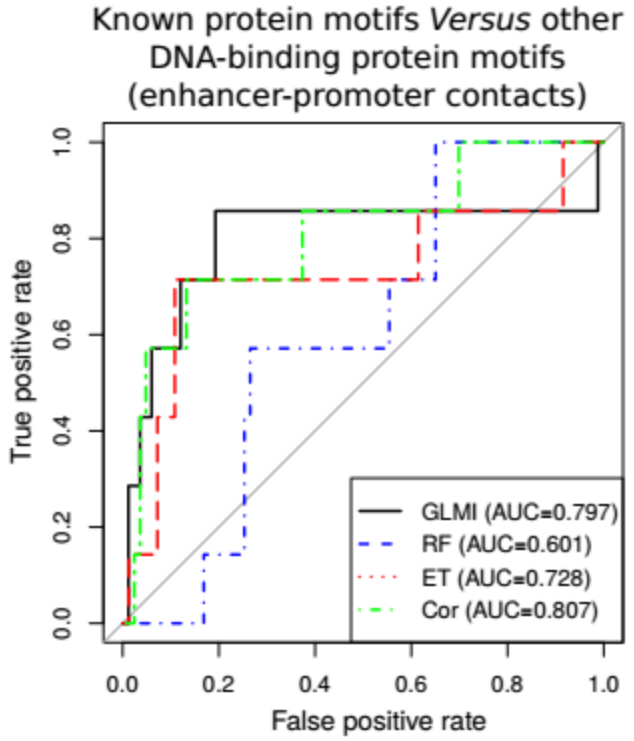
$$\begin{aligned}\log\left(\mathrm{E}[\mathbf{y}|\mathbf{X}]\right) &= \beta_0 + \beta_d\mathbf{d} + \boldsymbol{\beta}_B\mathbf{B} + \boldsymbol{\beta}_C\mathbf{C} + \beta_g\mathbf{g} \\ &= \beta_0 + \beta_d\mathbf{d} + \boldsymbol{\beta}_B\mathbf{B} + \beta_{m_i}\mathbf{m}_i + \beta_{m_j}\mathbf{m}_j + \beta_{m_k}\mathbf{m}_k + \beta_{m_{ik}}\mathbf{m}_{ik} + \beta_{m_{jk}}\mathbf{m}_{jk} \\ &+ \beta_{n_{ij}}\mathbf{n}_{ij} + \beta_{n_{jk}}\mathbf{n}_{jk} + \beta_{n_{ik}}\mathbf{n}_{ik} + \beta_{n_{kk}}\mathbf{n}_{kk} \\ &+ \beta_{n_{ij}\times m_k}\mathbf{n}_{ij} \times \mathbf{m}_k + \beta_{n_{kk}\times m_i}\mathbf{n}_{kk} \times \mathbf{m}_i + \beta_{n_{kk}\times m_j}\mathbf{n}_{kk} \times \mathbf{m}_j + \beta_{c_{ijk}}\mathbf{c}_{ijk}.\end{aligned}$$

Illustration du modèle



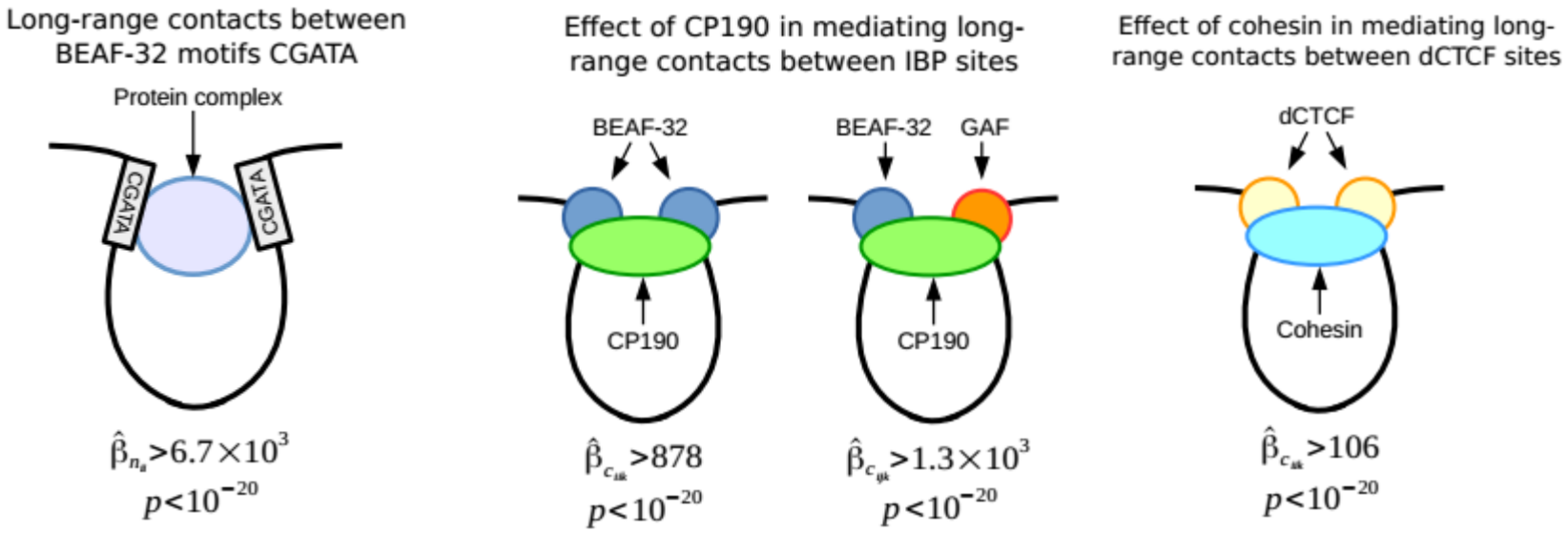
Comparaison sur données réelles de mouche





Le modèle linéaire généralisé avec interactions (GLMI) surpasse le test d'enrichissement (ET), les Random Forests (RF) et les corrélations (Cor) pour l'identification des facteurs connus.

Validations biologiques (mouche)



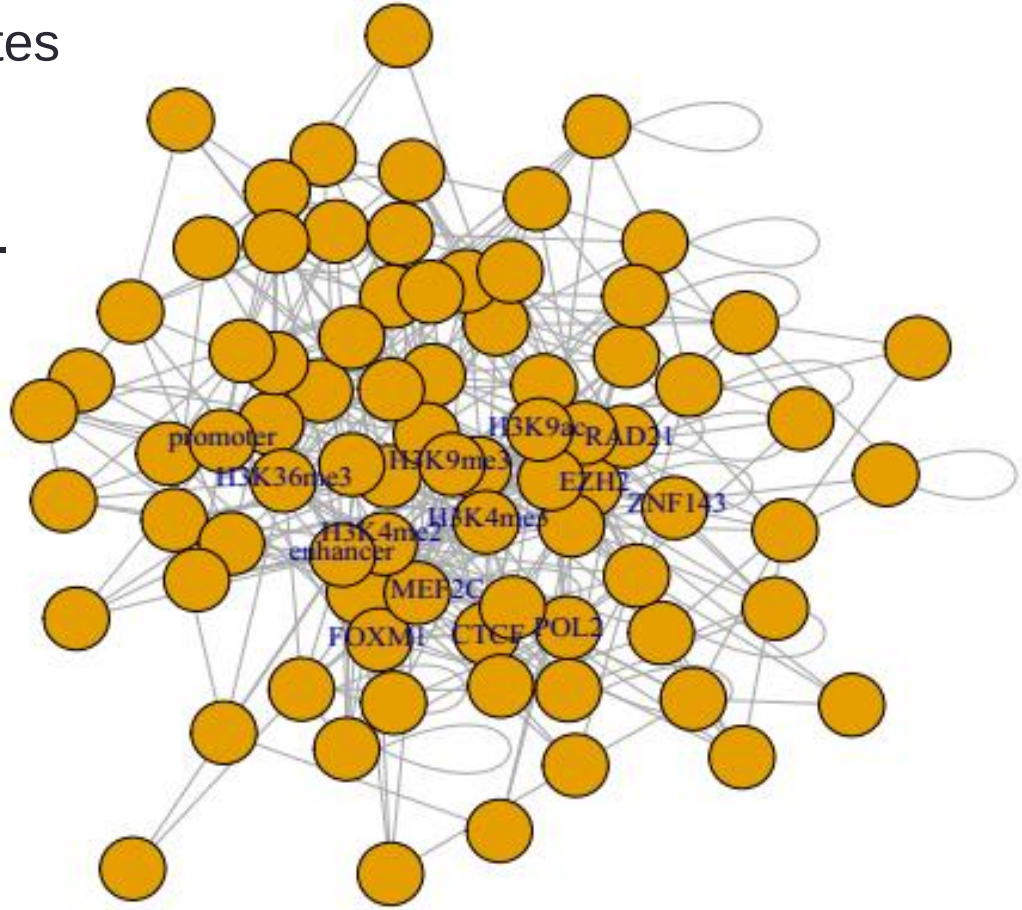
On a retrouvé de nombreuses interactions connues entre protéines architectes.

Réseau d'interaction longue-distance entre de multiples protéines et marques histones (homme)

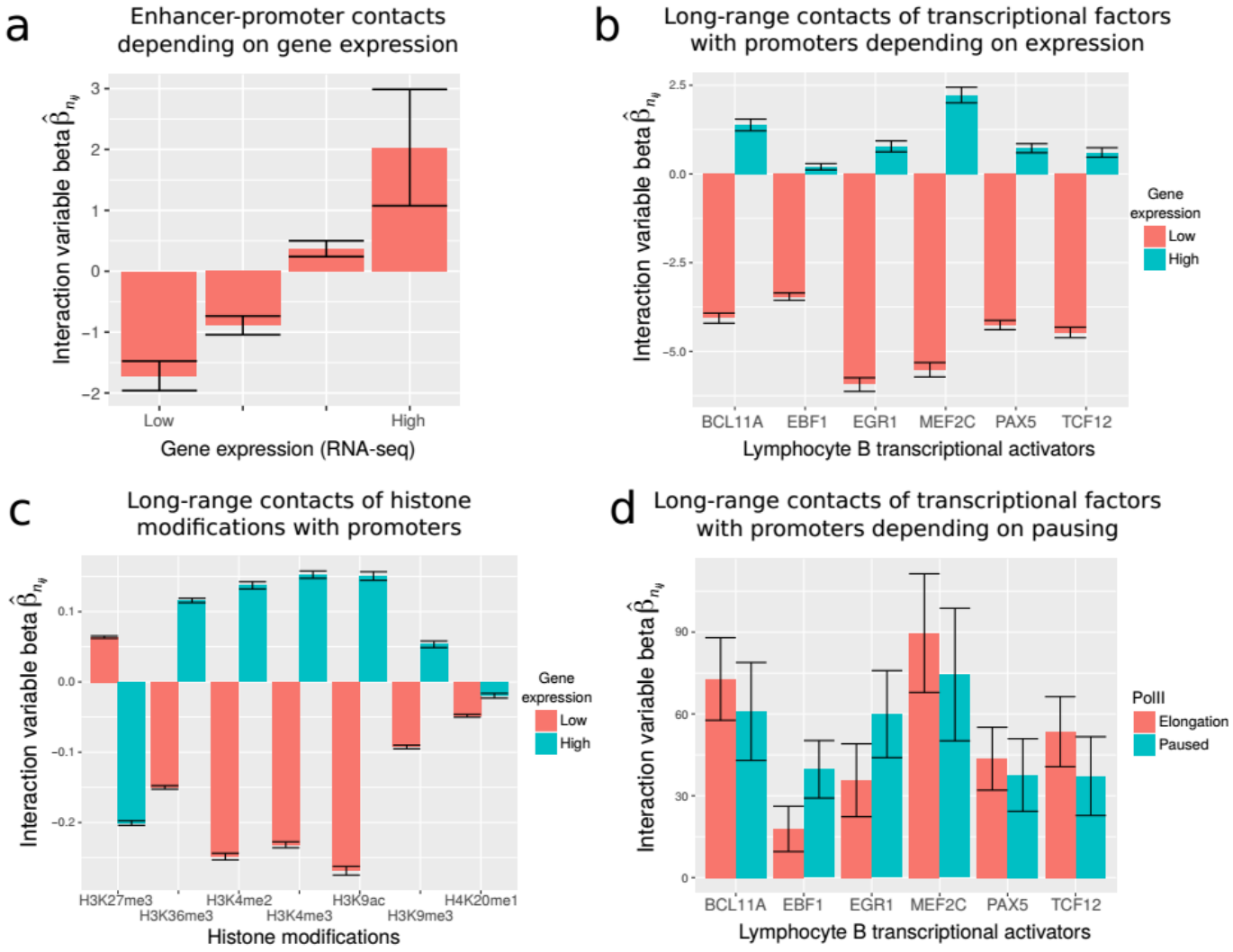
Ici modèle complet contenant toutes les interactions possibles (>3000 variables d'interaction) et apprentissage par Poisson Lasso.

Construction du graphe:

- Si $\beta_{nij} > 0$, alors arrête entre nœuds i et j;
- Si $\beta_{nii} > 0$, alors arrête entre nœud i et lui-même;



Interactions enhanceur-promoteur



CONCLUSION

Conclusion

- Le modèle linéaire généralisé avec interactions (GLMI) détecte mieux les interactions entre protéines que les approches courantes dont le test d'enrichissement, les random forests et la corrélation.

- Nous avons validé de nombreuses interactions connues dans la littérature.

- Le modèle a permis de montrer de nouveaux résultats:
  - La marque H3K27me3 réprime les gènes en 3D;
  - En fonction de la protéine impliquée, les contacts enhanceur-promoteur peuvent aussi associés à la polymérase en élongation.

BIBLIOGRAPHIE

Bibliographie

- 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.