

# Conséquences évolutives des réarrangements des paysages régulateurs de l'expression des gènes

Anamaria Necsulea

anamaria.necsulea@univ-lyon1.fr

## Introduction

La régulation de l'expression des gènes est un processus complexe, qui fait intervenir des éléments régulateurs proximaux (situés près des gènes, à des distances de l'ordre du millier de nucléotides) ou distaux (loin des gènes, à des distances qui peuvent aller jusqu'à plusieurs millions de nucléotides). Il s'agit de séquences d'ADN situées sur le même chromosome que les gènes qu'ils contrôlent ; on parle alors d'éléments régulateurs en *cis*. Pour réguler l'activité des gènes, ces séquences sont liées par des protéines appelées facteurs de transcription. La liaison entre facteurs de transcription et éléments *cis*-régulateurs peut avoir comme conséquence l'activation, l'amplification, l'inhibition ou la répression de l'expression du gène. L'élément régulateur le plus important est le promoteur du gène, une région de quelques milliers de nucléotides qui contient le site d'initiation de la transcription. Pour agir sur l'expression des gènes, les éléments régulateurs distaux interagissent avec le promoteur, grâce à la formation de boucles de chromatine qui amènent en proximité physique dans le noyau des régions séparées par de grandes distances dans le génome linéaire. La distribution génomique des éléments *cis*-régulateurs autour des gènes qu'ils contrôlent a été appelée "paysage régulateur" (Symmons et Spitz, 2013).

Intuitivement, on s'attendrait à ce que l'évolution de l'expression soit bien corrélée avec l'évolution des paysages régulateurs. Cependant, la plupart des études comparatives effectuées jusqu'à présent ont mis en évidence uniquement des corrélations légères (Berthelot *et al.*, 2018; Laverré *et al.*, 2022), voire pas de corrélation (Wong *et al.*, 2015), entre les vitesses d'évolution des deux processus. En général, l'expression des gènes évolue lentement (Brawand *et al.*, 2011), alors que les éléments régulateurs (notamment les éléments amplificateurs de l'expression) évoluent rapidement (Villar *et al.*, 2015). Une explication pour ce paradoxe pourrait résider dans l'existence d'une redondance entre éléments régulateurs. En effet, chaque gène est contrôlé par plusieurs éléments régulateurs, qui peuvent être actifs dans les mêmes types cellulaires, assurant ainsi une robustesse du patron d'expression devant les mutations qui peuvent inactiver chacun des éléments. Une autre explication pourrait être liée à un biais de sélection des gènes étudiés : pour évaluer l'évolution de l'expression des gènes, on analyse naturellement des familles de gènes orthologues, qui sont donc conservés et présents en une seule copie chez toutes les espèces analysées. Il est possible que certaines mutations qui affectent le paysage régulateur de l'expression des gènes soient tellement délétères qu'elles mènent à la pseudogenisation (ou perte) du gène affecté. Ces gènes-là sont *de facto* exclus des analyses, ce qui fait donc que l'on cache une partie importante des mutations régulatrices : celles qui ont le plus d'effet sur l'expression des gènes. Les grands réarrangements génomiques, qui perturbent complètement les paysages régulateurs (par exemple en éloignant les éléments régulateurs sur un autre chromosome, ou à une distance génomique qui est prohibitive pour les contacts de chromatine) pourraient faire partie de ces mutations à fort effet sur l'expression des gènes.

## Objectifs

Dans le cadre de ce projet, nous proposons de tester si les réarrangements génomiques qui affectent fortement les paysages régulateurs des gènes peuvent avoir comme conséquence la pseudogénéisation des gènes affectés. L'analyse se fera chez les vertébrés, d'abord sur un ensemble réduit d'espèces (humain, souris, poulet), pour lesquelles l'on dispose de données de séquençage de transcriptome (RNA-seq), ainsi que des données décrivant les contacts de chromatine entre promoteurs et éléments régulateurs (données Hi-C ou PCHi-C). Dans un premier temps, il faudra identifier des instances de pseudogénéisation de gènes, en utilisant les ressources disponibles dans la base de données Ensembl (Vilella *et al.*, 2009). Ensuite, il s'agira de mettre en lien ces instances de pseudogénéisation avec les réarrangements génomiques qui perturbent les paysages régulateurs, en utilisant les données traitées par Laverré *et al.* (2022). Pour déterminer la significativité statistique des résultats, nous pourrons réaliser des simulations des réarrangements des paysages régulateurs.

## Pré-requis

La mise en oeuvre de ce projet nécessite l'utilisation de la ligne de commande Unix (Bash), d'un langage de programmation tel que Python, ainsi que la mise en oeuvre de tests statistiques et la production de graphiques dans R.

## Techniques à acquérir pendant le stage

- Analyse des données de séquençage de transcriptome (RNA-seq).
- Analyse des données de capture de conformation de la chromatine centrées sur les promoteurs des gènes (PCHi-C).
- Analyse des familles de gènes homologues et de l'évolution des séquences codantes.
- Mise en place de simulations pour le test d'hypothèse.

## Bibliographie

- Berthelot, C., Villar, D., Horvath, J. E., Odom, D. T., and Flicek, P. 2018. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nature Ecology & Evolution*, 2(1): 152–163.
- Brawand, D., Soumillon, M., Necsulea, A., Julien, P., *et al.* 2011. The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369): 343–348.
- Laverré, A., Tannier, E., and Necsulea, A. 2022. Long-range promoter-enhancer contacts are conserved during evolution and contribute to gene expression robustness. *Genome Research*, 32(2): 280–296.
- Symmons, O. and Spitz, F. 2013. From remote enhancers to gene regulation: charting the genome's regulatory landscapes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 368(1620): 20120358.
- Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., *et al.* 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research*, 19(2): 327–335.
- Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., *et al.* 2015. Enhancer evolution across 20 mammalian species. *Cell*, 160(3): 554–566.

Wong, E. S., Thybert, D., Schmitt, B. M., Stefflova, K., *et al.* 2015. Decoupling of evolutionary changes in transcription factor binding and gene expression in mammals. *Genome Research*, 25(2): 167–178.