A genome-wide, machine learning-guided exploration of the cis-regulatory code involved in neuronal differentiation

Océane Cassan, Christophe Vroland, Julien Raynal, Kayoko Yasuzawa, Tsukasa Kouno, Jen-Chien Chang, Chung-Chau Hon, Jay W. Shin, Masaki Kato, Hazuki Takahashi, Takeya Kasukawa, Robert Lehmann, Vincenzo Lagani [many others from FANTOM6], Piero Carninci, Kévin Yauy, Chi Wai Yip, Laurent Bréhélin, Charles Lecellier

Montpellier Computational Regulatory Genomics group (ML4REGGEN)

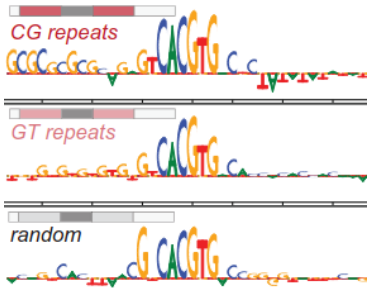
April 3, 2025

Learning the *cis*-regulatory code

What are the sequence features underlying the transcriptional activity of cis-regulatory elements (CREs) during dynamic processes like differentiation?

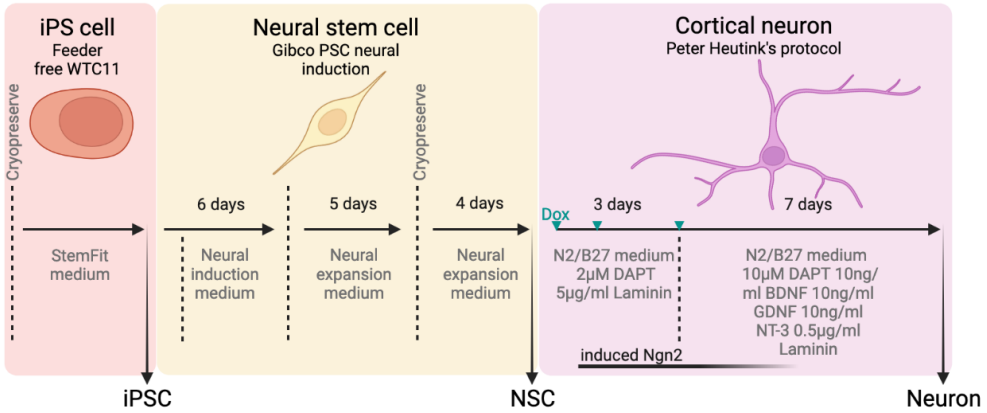
→ Relative impact of

- TF binding motifs (TFBMs)
- k-mer content and low complexity DNA?



[Horton et al., 2023]

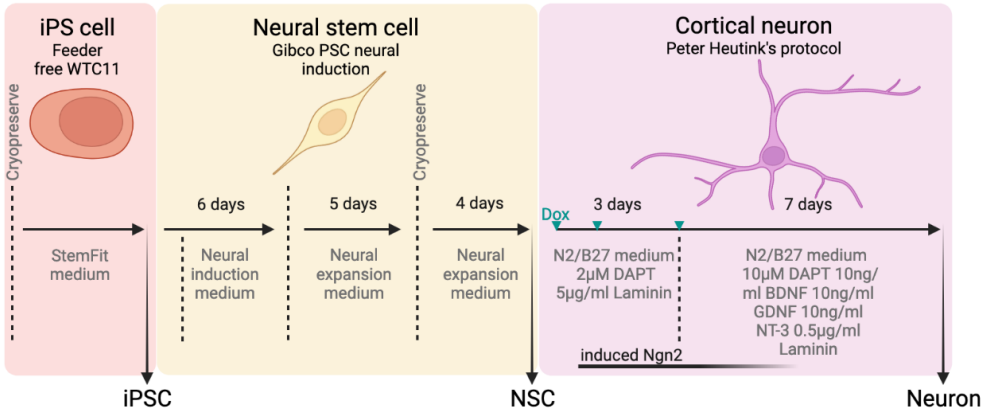
Case study: single cell dataset of neuronal differentiation



5' end **scRNA-Seq** data & **snATAC-Seq** data

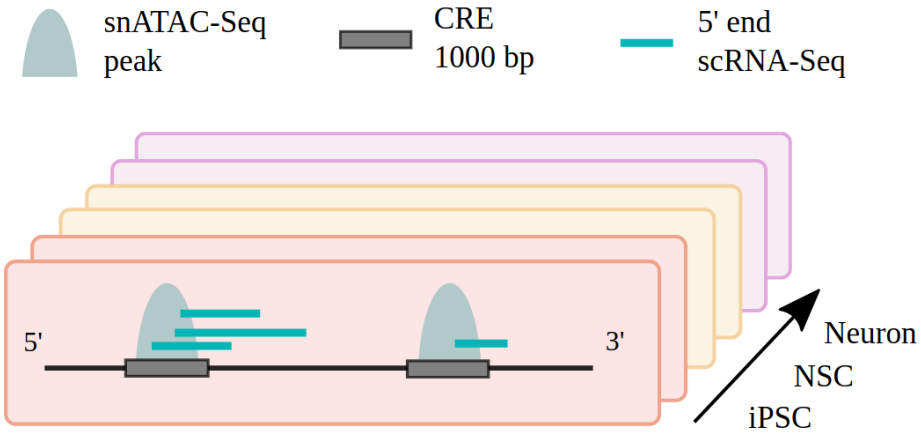
Data from Wallace Yip's lab: Kayoko Yasuzawa, Tsukasa Kouno, Jen-Chien Chang, Chung-Chau Hon, Jay W. Shin

Case study: single cell dataset of neuronal differentiation

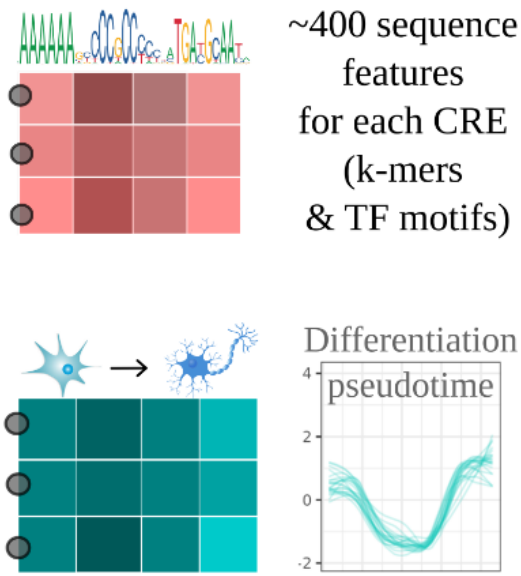


5' end **scRNA-Seq** data & **snATAC-Seq** data

Data from Wallace Yip's lab: Kayoko Yasuzawa, Tsukasa Kouno, Jen-Chien Chang, Chung-Chau Hon, Jay W. Shin

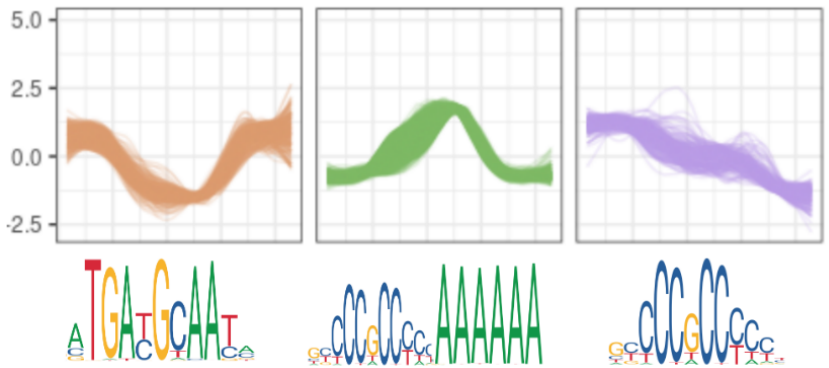


$N = 10912$ differentially expressed CREs along differentiation



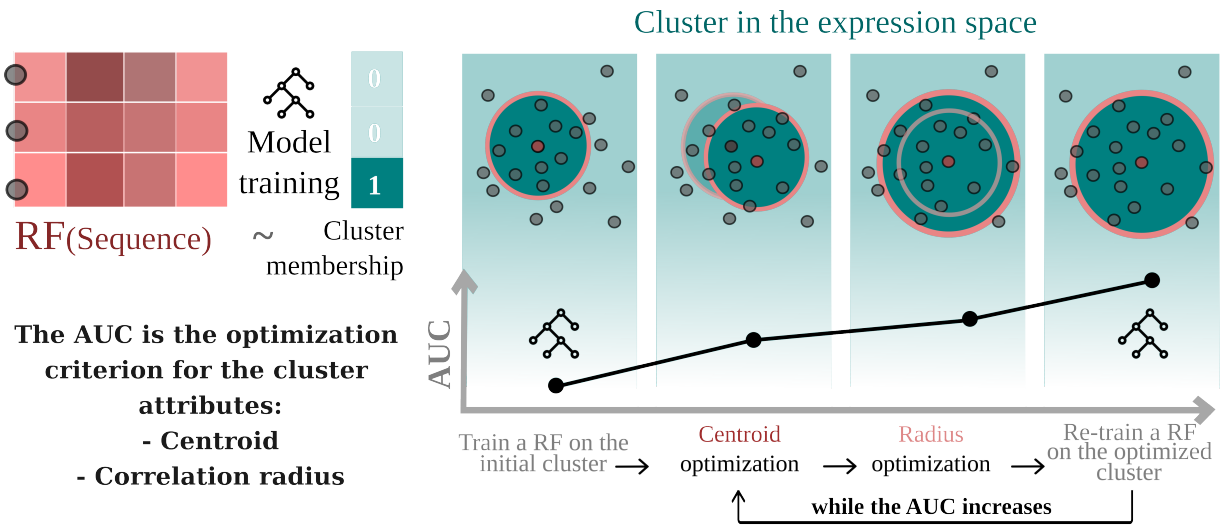
Learning the cis-regulatory grammar

How can we associate CRE sequence features to coordinated activity profiles?

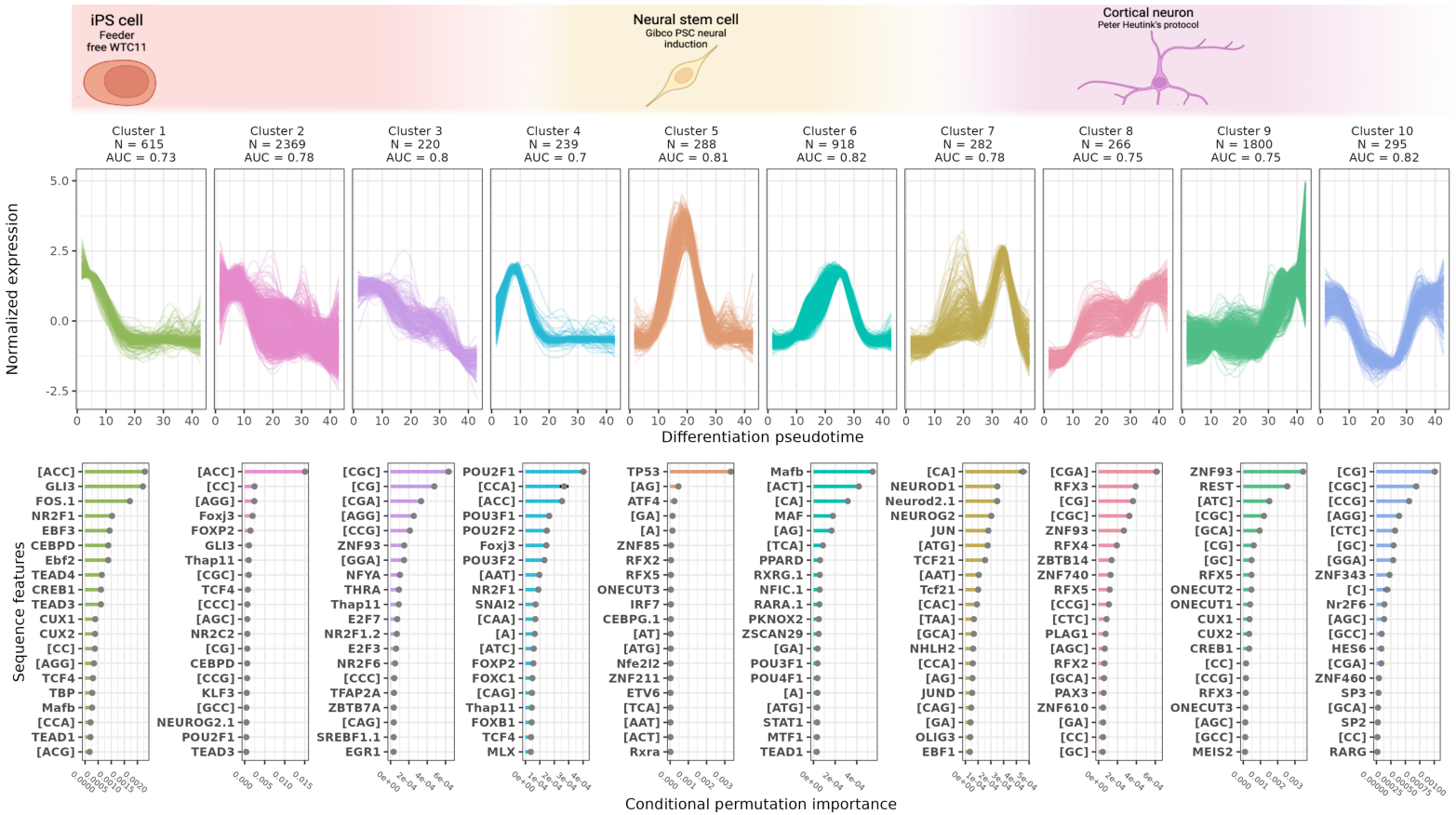


Supervised learning To Inform Clustering (STOIC)

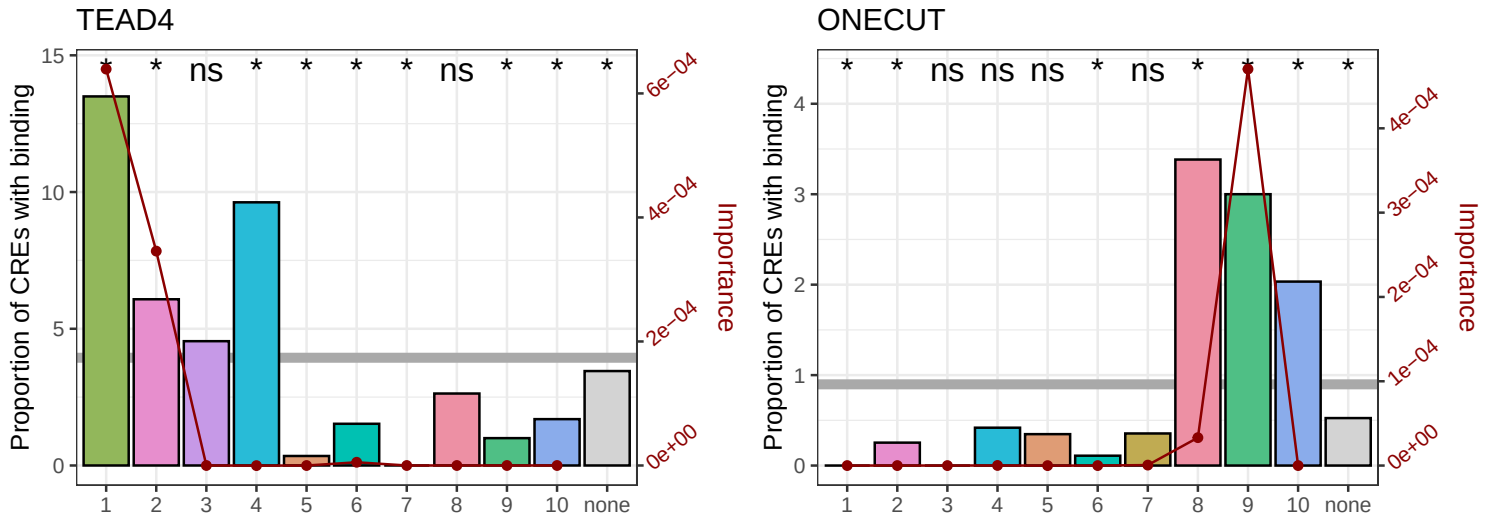
Co-expression clusters are learned to maximize the **AUC of a model using sequence features** to predict cluster expression membership



Diverse activity profiles are identified by STOIC

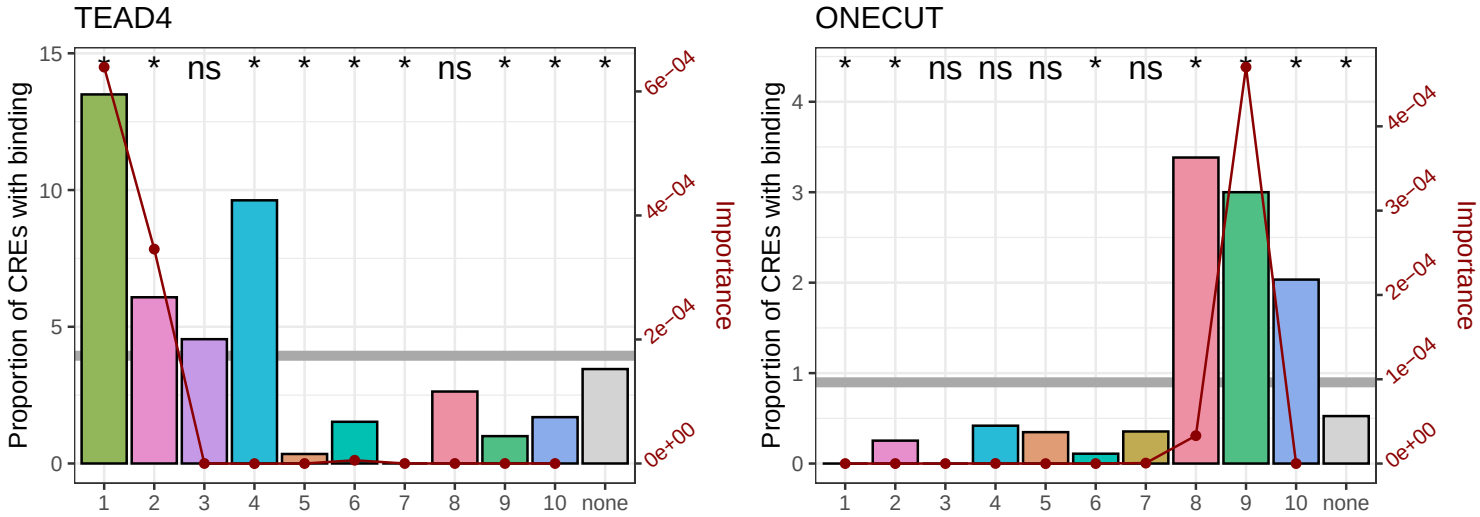


Important features supported by CHIP-Seq and functional databases

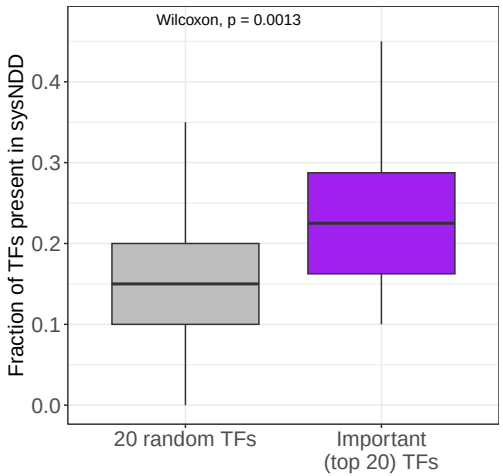


Data from Wallace Yip's lab

Important features supported by CHIP-Seq and functional databases



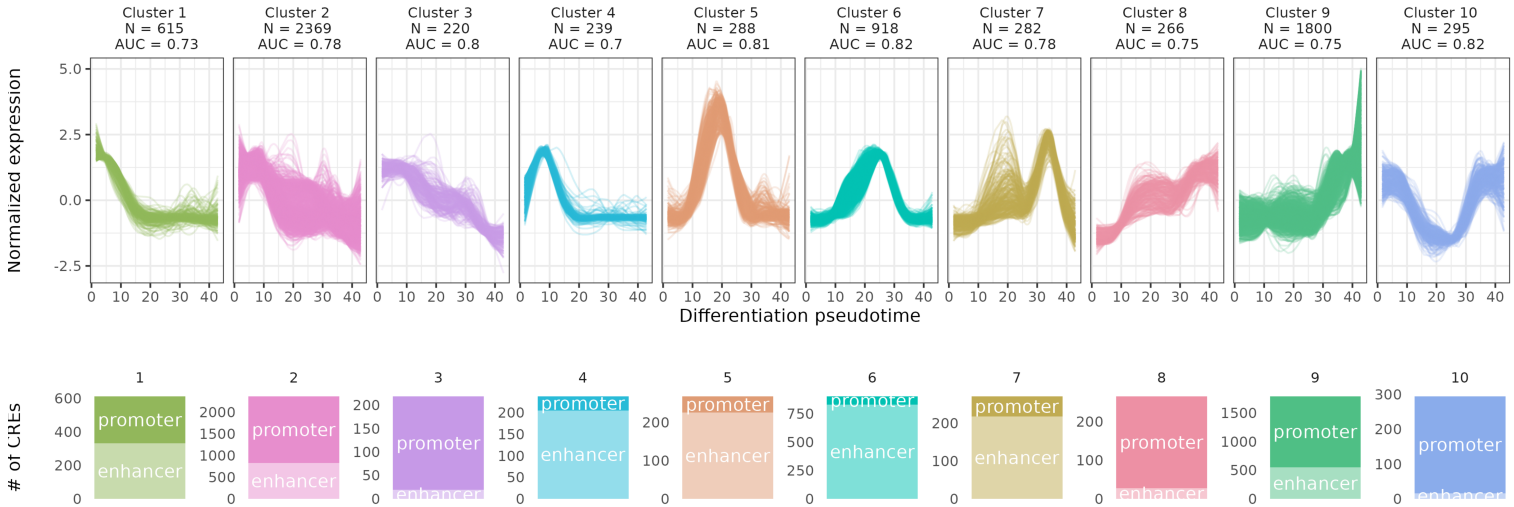
Data from Wallace Yip's lab

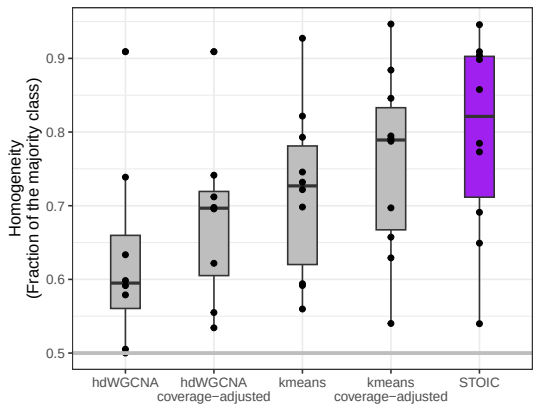


Important TFs are enriched in neurodevelopmental disorders related genes. Source: **sysNDD** database `https://sysndd.dbmr.unibe.ch`, a curated database of gene disease relationships in neurodevelopmental disorders. Analysis in collab. with Kevin Yauy

Homogeneous epigenetic marks within inferred clusters

Stoic clusters are very often either **enhancer-rich or promoter-rich**





Annotations derived from chromHMM applied to matched **CUT&Tag data** from Wallace Yip's lab

Perspectives

- STOIC's features are an interesting basis to **investigate the cis-regulatory code** (interplay between kmers and TFBMs) underlying coordinated coordinated regulatory processes
- Cross compare STOIC's features with patients SNPs to **help diagnose**

Stoic R package

Stoic's methodology is available as an R package. The machine-learning guided approach developped in Stoic is applicable to any problem where the clustering of some measurments can be guided by a second matched dataset.

```
library(remotes) # remotes should be installed if it is not
install_gitlab("oceane.cssn/stoic")
```

Acknowledgments

- The **ML4REGGEN** team

People

- Quentin Bouvier, PhD student, IGMM
- Elliot Butz, PhD student, LIRMM
- **Laurent Bréhélin**, CR CNRS, LIRMM
- Océane Cassan, Post-doc, LIRMM
- **Sophie Lèbre**, MCF Univ. Paul-Valery, IMAG & LIRMM
- **Charles Lecellier**, DR CNRS, IGMM & LIRMM
- Julien Raynal, PhD student, IGMM & LIRMM
- Mathilde Robin, Engineer, ICM & LIRMM
- Diego Tosi, MD PhD, ICM
- Christophe Vroland, Post-doc, IGMM & LIRMM
- Kevin Yauy, MD, PhD, Univ. Montpellier & CHU

igmm INSTITUT DE GÉNÉTIQUE MOLÉCULAIRE DE MONTPELLIER

IMAG INSTITUT MONTPELLIERAIN ALEXANDER GROTHENDIECK

LIRMM

- Our **FANTOM6** collaborators:

Wallace Yip for all the data and help, Kayoko Yasuzawa, Tsukasa Kouno, Jen-Chien Chang, Chung-Chau Hon, Jay W. Shin, Robert Lehmann, Vincenzo Lagani

References I

- Horton et al. (2023).
  Short tandem repeats bind transcription factors to tune eukaryotic gene expression.
  Science, 381(6664):eadd1250.