# ANALYSIS OF CELL FATE DYNAMICS VIA SINGLE CELL RNA-SEQ: ROLE OF LONG NON-CODING RNAS

**Supervisor**: **Olivier Martin** **Co-supervisor: Thomas Blein**
mail: olivier.c.martin@inrae.fr mail: thomas.blein@cnrs.fr
Institute of Plant Sciences – Paris-Saclay, Bâtiment 630, rue de Noetzlin, 91405 – Orsay, France
Web: Team REGARN: Regulatory non-coding RNAs in root plasticity

## INTERNSHIP DESCRIPTION

Single-cell transcriptomics [1] has generated a revolution in our knowledge of cell types and cell fates. In particular, computational teams working in single-cell RNA-seq (scRNA-seq) data analysis have developed intuitive software tools [2,3] to automate clustering, dimensional reduction, and heterogeneous data merging. Thanks to these tools, multiple groups have constructed expression atlases for different species [4,5] and the field is now moving very fast because it provides such powerful and detailed quantification. In the corresponding analyses, it is now possible to infer developmental "trajectories" from the gene expression patterns, for instance when going from pluripotent to terminally differentiated states. As of this summer, these tools have been integrated into a local in-house pipeline produced during a Ph.D. project. In this personalized software, the parameters have been adjusted, ensuring that the different steps lead to reliable predictions. The objective of this M2 project is to use this pipeline to identify **long non-coding RNAs** important for root development in the model plant *Arabidopsis thaliana*.

During this internship you will mine a dozen publicly available single-cell and single-nucleus RNA-seq datasets, all produced for studying Arabidopsis root development. You will compare two methodologies: (1) performing a separate analysis for each dataset followed by a merging of the different predictions to form a consensus prediction, and (2) merging the different datasets and running the pipeline thereon to obtain a single prediction. At stake are the relevance and challenges of merging datasets produced using different protocols and technologies, each having its own – generally unknown – biases. In particular, although it is quite preliminary, it seems that single-cell (sc) and single nucleus (sn) approaches each have their advantages and drawbacks so that it should be possible to build on their complementarities. We thus aim to provide new ways to perform data integration to enhance these types of analyses. Furthermore, there is a significant difficulty compared to previous work: the expression levels of long non-coding RNAs are far lower than those of coding genes. Only with multi-datasets will these low signal-to-noise ratios be overcome.

The exploitation of these sc and sn RNA-seq datasets will provide a robust inference of developmental trajectories that will realize Waddington's "epigenetic landscapes" [6] with

associated branching processes. The trajectories for both coding and non-coding genes will be compared to infer candidate regulatory drivers and interactions (e.g., which non-coding might regulate which coding genes). Furthermore, our preliminary results show that non-coding genes will allow us to identify cellular sub-types as yet unresolved by coding genes. The predictions you produce will be curated in collaboration with the wet lab biologists of the host team to unravel the genetic and molecular processes driving root development in *Arabidopsis thaliana*.

This work will form a stepping stone for integrating more heterogeneous datasets, a topic that will be considered in a potential follow-up doctoral work. The goal there will be to integrate single-cell ATAC-seq datasets with the transcriptomic ones studied in this internship project, and to also perform this type of integration across different organisms.

## TECHNIQUES USED DURING THE INTERNSHIP

The computational work will be done at least partly in the language R to exploit the Seurat [2] and Monocle [3] packages that are easy to use and allow the identification of cell types and pseudo-time trajectories. The work will require running the current pipeline, testing different parameter and algorithmic choices, quantifying inference reliability and exchanging with wet-lab biologists to confirm predictions for the regulatory interactions.

## REFERENCES

[1] Chen et al. (2019), Single-Cell RNA-Seq Technologies and Related Computational Data Analysis: https://doi.org/10.3389/fgene.2019.00317

[2] Seurat computational tools: https://satijalab.org/seurat/index.html

[3] Monocle computational tools: http://cole-trapnell-lab.github.io/monocle-release/

[4] Mouse Organogenesis Cell Atlas. https://oncoscape.v3.sttrcancer.org/atlas.gs.washington.edu.mouse.rna/landing

[5] Plant sc-Atlas, https://bioit3.irc.ugent.be/plant-sc-atlas/root

[6] Waddington, C.H. *The Strategy of the Genes*. London: Geo Allen & Unwin, 1957