GRN Inference Analysis: Guiding Questions

Erwan Cruché et Sergio Peignier

February 2026

Understanding the Biological Context

- What is the main objective of the Potier et al. paper? What biological question are the authors trying to answer?
- What assumptions do the authors make about gene regulation during *Drosophila* eye development?
- What experimental or computational approach did they use to perturb the transcriptome?
- How do they define "regulatory relationships" in their analysis?
- What is a gene regulatory network (GRN), and why is it important in developmental biology?
- Which organism, tissue, and developmental stages are studied — and why are they relevant?
- What prior knowledge or databases do they rely on (e.g., known TFs, motifs, gene annotations)?
- How do they validate or interpret their inferred regulatory networks?
- What are the key limitations or challenges they acknowledge in their approach?
- How does *Drosophila* eye development relate to the study of GRNs?
- What biological insights can we gain from mapping transcriptional relationships during eye development?

Data Preparation & Exploration

Gene expression data

- What format is the gene expression data in, and how should you load it for analysis?

- After loading the expression matrix, what checks can you perform to verify the data was imported correctly?
- How do you inspect the dimensions, data types, and structure of your expression matrix?
- What should you check regarding gene names and sample identifiers (e.g., are they unique, properly formatted)?
- Are there missing values or NaN entries in the dataset?
- What is the range of expression values, and does it suggest about the data?
- What filtering/normalization steps should be performed before GRN inference?
- Are there any batch effects or technical artifacts you need to account for?

Finding *Drosophila melanogaster* TF Lists for pySCENIC

- Where is the AnimalTFDB located, and what type of data does it provide?
- How do you search for transcription factors specific to *Drosophila melanogaster*?
- What file format or identifier system (e.g., gene symbols, FlyBase IDs) does AnimalTFDB use?
- How do you export or download the TF list for use in pySCENIC?
- Does pySCENIC require any preprocessing of the TF list (e.g., renaming, filtering)?
- How do you verify that the TFs you downloaded are compatible with the rest of the data?

pySCENIC pipeline

Understanding pySCENIC

- What is pySCENIC, and what is its primary objective in GRN inference?
- What are the key underlying assumptions of the pySCENIC pipeline?
- How does pySCENIC combine co-expression analysis with motif-based regulatory information?
- What are the main advantages of pySCENIC compared to other GRN inference tools?
- What are potential limitations or weaknesses of the pySCENIC approach?

- How does pySCENIC handle indirect vs. direct regulatory relationships?
- What type of data does pySCENIC require?
- How sensitive is pySCENIC to parameters, and which ones are most critical to tune?
- What is the computational complexity and runtime expectation for a typical analysis?
- Are there example scripts or tutorials in the pySCENIC repository that show the full workflow?
- Where do you plan to find solutions for the errors you might you encounter?

Step 1: Co-expression analysis

Running Arboreto

- Where can you find the pySCENIC documentation, and what is the main function for co-expression?
- What input files do you need (expression matrix format, gene names)?
- How do you import and use the GRNBoost2 or GENIE3 functions?
- What parameters should you consider when running GRNBoost2 or GENIE3, and where are they documented? Can you change them?
- What output file format does GRNBoost2 or GENIE3 produce, and how do you interpret it?
- Compare the scores obtained by GRNBoost2 and GENIE3, what do you conclude?

Converting Network to Regulon Modules

- What is the purpose of the `modules_from_adjacencies` function? What information does the output list of modules contain?
- What is a regulon module, and how does it differ from a raw regulatory relationship?
- What does the pruned network represent, and what format should it be in?
- Why do you need to transpose the expression matrix (`expr.T`) when creating modules?
- How do you inspect or iterate through the modules to understand their structure? What does a regulon contain in pySCENIC?

Step 2: Motif Pruning

Loading Ranking Databases

- What is a FeatherRankingDatabase, and why is it used in pySCENIC? What file format are the ranking databases in, and what information do they contain?
- Where do you find the ranking database files for *Drosophila melanogaster*?
- Why must you call `db.load_full()` on each database before pruning?
- What happens if you skip the `load_full()` step?

Pruning the Network

- What is the purpose of the `prune2df()` function?
- What are the three main inputs to `prune2df()`, and what does each represent?
- What does the pruning step filter for?
- What is the output format of `prune2df()`, and what information does it contain?
- How do you convert the pruned dataframe to regulon objects using `df2regulons()`?
- How many regulons do you obtain after pruning, and what does this number tell you?

Step 3: AUCell Scoring

- What is AUCell, and why is it used after motif pruning?
- Which pySCENIC module contains the AUCell function?
- What input files does AUCell require?
- What is the output format, and how do you interpret the activity matrix?

Analysis & Interpretation

- Which transcription factors emerge as key regulators in eye development?
- How do the inferred regulons compare across different developmental stages or cell types?
- How do you extract and list the most significant regulators from the AUCell output?

- What tools or databases can you use to cross-reference your findings with known eye development regulators?
- What biological validation can you perform on your results?
- How do your findings relate to known developmental pathways?