Analysis of HVG use in the ScReNI Pipeline

Mihnea-Matei Gusu - mgusu@tudelft.nl

EEMCS - TUDelft - CSE3000 Research Project

Objectives & Research Question

Analyze the use of Highly Variable Gene (HVG) use in Single-cell Regulatory Network Inference (ScReNI) and answer these questions:

- **Does a cell-specific HVG produce more accurate (precision/recall on Chip-seq data) results when applied on ScReNI?**
- **Does the cell-specific HVG produce more biologically significant results?**

Introduction - What are HVGs?

- The whole genome is too large to be used for ScReNI.
- Some genes stay constant while other vary wildly.
- We are interested in the ones that vary, as those are the ones useful for regulation.
- So ScReNI chooses the top K highly variable genes (HVGs), which we call the global approach
- But what if the genes important to one cell-type get over-varied by the genes of another cell-type and end up being missed?
- Would it be better to select top K genes from each individual cell-type?
- These 2 HVG selection approaches are compared in this paper

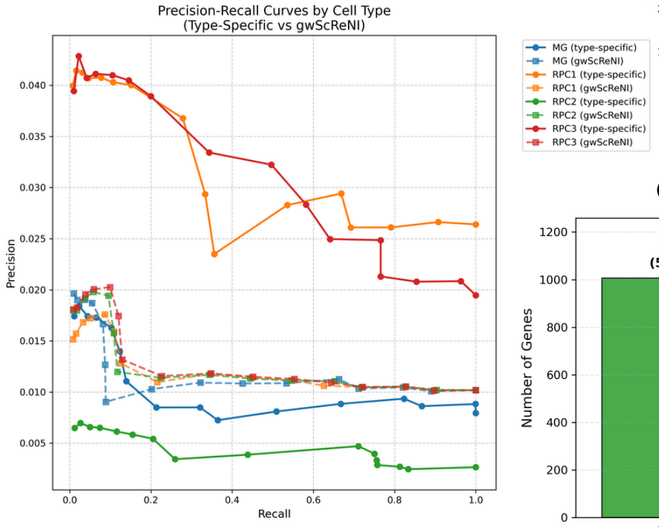
Methods and Comparison Metrics

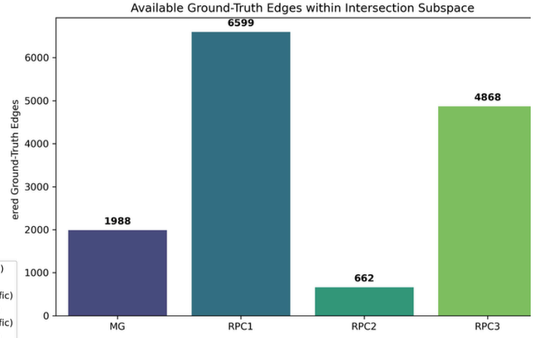
After converting the original ScReNI code from R into Python, we compare the methods by these metrics:

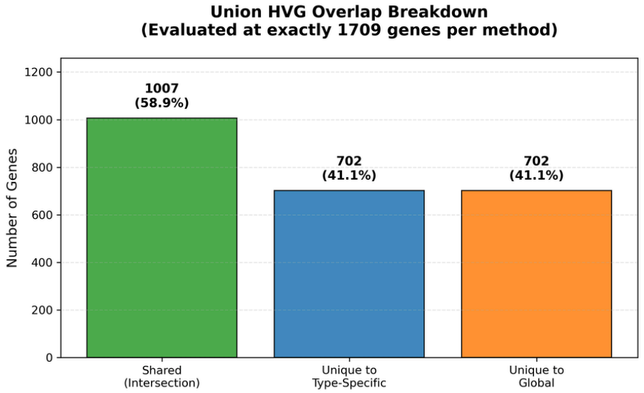
- Precision and Recall on ChIP-seq Atlas
- Modularity of the resulting GRNs and gene set enrichment on the modules
- Main regulator and plots of the GRNs

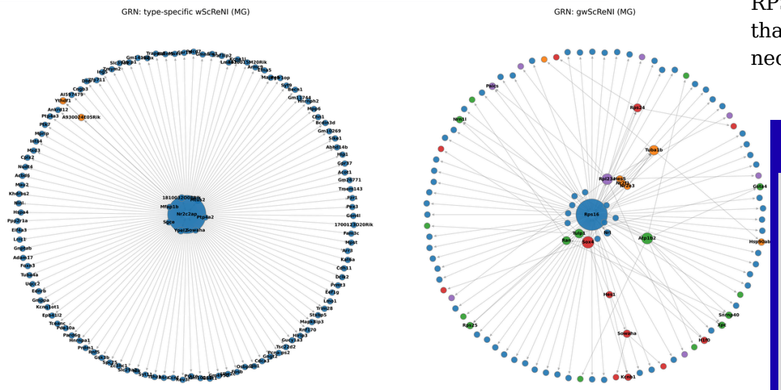
Main Results

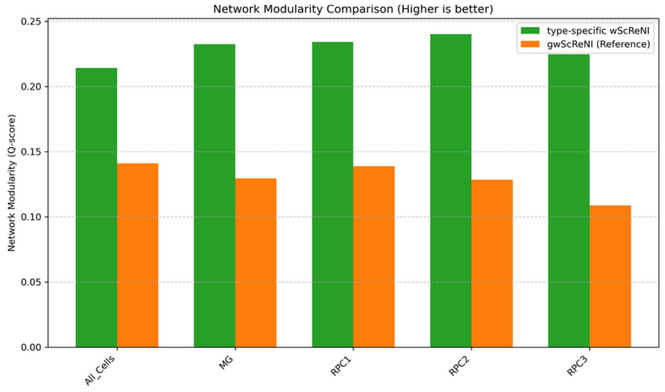
The type-specific selection had better results on precision and recall, modularity and main regulators. The gene set enrichment shows no conclusive results for the cell-specific, and only general pathways for the global approach.











Discussion

- Precision almost doubles for cell types RPC1 and RPC3. We can also see more different shapes in the cell types, rather than the same shape for all.
- Different levels of precision/recall can be attributed to the amount of edges that exist in ChIP-seq between the HVGs selected.
- Modularity is almost double the amount of the global selection.
- Gene set enrichment was run on the modules and returned inconclusive results for type-specific.
- On global Cytoplasmic Ribosomal Proteins; Dopaminergic Neurogenesis and Neural Crest Differentiation
- on the other hand, for global the largest regulator in the GRN plots was RPS16

RPS16 is not a regulator, It is a "housekeeping gene" that is expressed in a lot of progenitor cells, as it necessary for the propagation of the cell.

Conclusion

The global approach fixates on "housekeeping" genes, genes whose expression is large, but play no roles in regulation of other genes. Because the algorithm is run on early development cells, these genes have a high expression in some cells, that as the cell-cycle continues drops sharply, making them prime targets for the global HVG. This does not happen when taking the type-specific HVGs as the gene remains largely constant in one cell-type.

TU Delft