Analysis of HVG use in the ScReNI pipeline

A comparison of global and cell-type specific HVG selection

Mihnea-Matei Gusu
Supervisors: Marcel Reinders, Bram Pronk, Timo Verlaan
EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Name of the student: Mihnea-Matei Gusu
Final project course: CSE3000 Research Project
Thesis committee: Sicco Verwer, Marcel Reinders, Timo Verlaan, Bram Pronk
21st of June, 2026

Abstract

ScReNI[21] is a recently developed algorithm that aims to infer the gene regulatory network (GRN) of single cells based on both single-cell RNA sequence (scRNA-seq) and single-cell ATAC sequence (scATAC-seq) data. Because of its novelty, not much is known about its intricacies, which we aim to highlight in this paper. Specifically, this is a comparison of the highly variable gene (HVG) selection of ScReNI and a newly proposed selection approach: the type-specific HVG selection. Instead of taking the top HVGs by decreasing value from the entire dataset, we propose taking the top HVGs present in each cell type and inferring the GRN of each cell only with the genes specific to its cell type. The comparison was made across a number of metrics, both structural and biological, and the type-specific selection has produced better results than the original approach overall.

1 Introduction

In the interest of curing diseases previously thought to be incurable, scientists have in recent times begun looking at ways to better understand how the human body works at a deeper and deeper level. One such way is gene regulatory networks (GRN) which are graphs that show how exactly genes influence each other. For example, if one gene is expressed does that mean that another gene is inhibited or does the presence of a gene cause another gene to also be expressed? Although traditionally, GRNs have been computed for a whole tissue sample containing millions of cells, recent developments have made single cell GRNs possible through the sequencing of RNA of single cells. This would enable the observation of how exactly a biological process or disease influences the gene expression of a single affected cell, potentially making a deeper insight possible into how so-called "incurable" diseases such as Alzheimer's Disease [5] (AD) change the body at a cellular level.

A recent algorithm that shows great promise of computing these single cell GRNs is ScReNI (single cell regulatory network inference)[21] introduced by Xu et al. By leveraging both single-cell RNA and single-cell ATAC (assay for transposase-accessible chromatin; which part of the DNA is accessible) data and by being inherently robust against sparsity, it produces better results for single cell GRNs than other present day alternatives such as CSN[7], LIONESS[13], scGeneRAI[12], CeSpGRN[25] or LINGER[23].

Also, due to the large number of genes in the entire genome (approximately 36,000) it would be computationally unfeasible to run ScReNI on the whole genome. Given that genes that remain constant throughout the data do not provide much biologically relevant information for this context, the assumption was made that taking the top K highly variable genes (HVGs) will provide a representative subset of the original data. As it turns out, Highly Variable Gene selection [18] is a classical way to handle the large and sparse amount of data collected by the single cell gene sequencing algorithm. This technique is used in the ScReNI pipeline by taking either the top 500 or the top 2000 genes from the overall set of genes (hereafter referred to as "global"). There can be issues that arise from this method, as some cell types could be underrepresented in the HVG selection. This would mean that there can be cell types whose most important genes do not vary enough for them to be in the HVGs and therefore not be included in the GRNs inferred.

This paper aims to verify if that is indeed the best way of using the global technique or if there are other, better, ways of representing the data with a different HVG approach. In particular, a cell-type specific HVG selection (taking the top K highly variable genes of every cell type) will be compared to the original global HVG approach. This paper will present a number of comparisons of the output of ScReNI with a global and a cell-type specific HVG selection approach. The main comparison point will be the wScReNI algorithm which shows the best results in the original ScReNI paper[21]. It will from now on will be referred to as type-specific wScReNI for the one running with type-specific HVG selection and gwScReNI for the one running with global HVG selection.

2 Methodology

2.1 Background

ScReNI uses the paired single-cell RNA-seq and ATAC-seq data to compute the k nearest neighbors (weighted nearest neighbors for wScReNI) of a cell in terms of gene expression using dimensionality reduction. It then feeds the neighbor's gene sequences into a random forest model that is trained to determine importance of gene to gene edges with the formula:

$$X^q = f_q(X^{-q}, Y^q) + \epsilon_q$$

Based on the importance ($z_{i,j}$) and the peaks in the ATAC data (p_{il}) the weight is calculated using the formula:

$$w_{i,j} = z_{i,j} + \sum_{l} p_{i,l} I_{i,j}$$

where $I_{i,j}$ is 1 if gene j binds to peak l of gene i. A weight threshold is then chosen and any edge with a weight over that threshold is considered part of the resulting GRN (Figure 8 from the ScReNI[21] paper is a visual explanation in the appendix).

2.2 Data

The ScReNI pipeline was run on four cell types of mouse retinal development data as in the original ScReNI paper[21]: RPC1, early progenitors, whose function is to rapidly multiply and expand the size of the developing retina; RPC2, an intermediate state where progenitors commit to becoming early-born retinal neurons, such as retinal ganglion or amacrine cells; RPC3, late-stage progenitors that appear toward the end of retinal development to become cells like rod photoreceptors, bipolar cells, and Müller glia; and MG (Müller Glia), the primary support cells of the mature retina that maintain structural and metabolic homeostasis, emerging late in development from retinal progenitors.

2.3 Choosing HVGs

The global HVG approach was switched to the cell-type specific one. This approach takes the genes present in the expression of cells of each specific cell-type and selects the top K (500 in this case) genes that vary the most in value. In this way, the rest of the pipeline remains largely unchanged, allowing a good comparison between the different methods of choosing HVGs. The global HVG selection was also changed to select the top N HVGs where N is equal to the number of HVGs in the union of the cell-type specific ones, in order to maintain information parity between the two approaches. In this way, both approaches will have, in total, access to the same number of genes and the only difference between them will be how they are selected.

2.4 Inferring the GRNs

The HVGs are then processed in accordance with the ScReNI pipeline. When inferring the GRNs, the pipeline diverges from the original implementation by having the random forest use only the HVGs of the cell type when inferring the GRN of a cell's type. This should have the effect of decreasing the computation time, as the random forest does not take into account HVGs that do not belong to the 500 HVGs of that cells cell-type. In this paper weighted ScReNI is the main focus of comparison, so from now on the cell-type specific wScReNI will be referred to as type-specific wScReNI and the global HVG (original) ScReNI will be referred to as gwScReNI.

2.5 Precision/Recall on ChIP-seq

In order for the inferred GRNs to be verified as either accurate or inaccurate, a ground-truth is needed, and the authors of ScReNI have decided to use ChIP-seq Atlas[26]. ChIP-seq Atlas is a database, which is a widely used collection of mouse retinal data from chromatin immunoprecipitation sequencing experiments. The data is generated by separating the DNA from the attached proteins and observing the interactions with accessible chromatin. This creates peaks in the ATAC-seq data which show where the strands of RNA bind to the chromatin giving us "edges". These edges then act as a ground truth for the GRNs of single cells that ScReNI outputs. The ChIP-seq Atlas database was filtered to only include edges between genes present in the intersection of the HVGs used by the two selection methods, so genes present in both the cell-type HVGs and the global HVGs. This would mean that no approach is evaluated against edges that it would have no way of inferring, as the genes in those edges are not present in the HVGs available to that approach, ensuring a fair comparison between them.

The metrics used to evaluate the GRNs on ChIP-seq are precision and recall, metrics also used by Xu et al. in the evaluation of ScReNI[21]. Precision is defined as:

$$TP/(TP + FP)$$

where TP is the number of true positives and FP is the number of false positives and recall is defined as

$$TP/(TP + FN)$$

where TP and FP are as above and FN is the number of false negatives. Because of this, precision penalizes the algorithm for its false positives and recall penalizes the algorithm for its false negatives. Given that the random forest outputs a weight for every edge possible in the gene-space, and that said gene space is variable (gwScReNI has 1709x1709 and type-specific only 500x500) the top 1% of possible edges (sorted by decreasing weight) are taken into consideration as part of the GRN of a cell and the precision/recall are calculated on this threshold.

2.6 Modularity

In order to derive the overall shape of the GRNs and to derive biological meaning from the inferred data, the modularity of the resulting GRNs at 1% threshold was calculated using Louvain community detection. This algorithm maximizes the Modularity Q metric, which is calculated according to the following formula:

$$Q = \frac{1}{2m} \sum_{i,j} (A_{i,j} - k_i k_j / 2m) \delta(c_i, c_j)$$

where $2m$ is sum of the degree of all genes; $A_{i,j}$ is 1 if there is an edge between gene i and gene j and 0 otherwise; k_i and k_j are the degree of the genes i and j; and $\delta(c_i, c_j)$ is the Kronecker delta of the communities of genes i and j, meaning it is 1 if $c_i = c_j$ and 0 otherwise. A higher Q means the communities formed have a higher number of edges between the genes in the same community and a lower number of edges between genes in different communities. High modularity in GRNs is also known to have strong ties to biological function and phenotypic robustness as presented by Kadelka et al. in their paper[11]. This modularity was run on a mean of all GRNs in one cell-type from both the type-specific wScReNI and gwScReNI, as this mean was considered sufficient to take into account the possible outliers of individual single-cell GRNs.

2.7 Gene set enrichment

For a proof of biological relevance, gene set enrichment was conducted on the modules found with Louvain community detection. Gene set enrichment is a technique which allows us to derive biological information from large sets of genes. By providing a background (all the genes that the algorithm had the option to select from) and a subset of those genes (in this scenario, our modules), gene set enrichment checks how many of the genes in the subset have been attributed in the queried database to a specific function or cell-type. In this way, it has become possible for the genes in the different modules to now have biological significance attributed to them. As background for the type specific modules the set of 500 HVGs of that cell-type was used, while for the global HVGs the background was the entire set of 1709 HVGs. The enrichment was done on the *WikiPathways 2024 Mouse* database[3] as this database aligns with our ChIP-seq mouse data and should show the biological pathways associated with the modules. The output of gene set enrichment comes in the form of a table containing the rows: Term (here referring to the found Pathway); Adjusted P-value; Odds Ratio; Genes. The adjusted P-value is calculated based on the following formula:

$$P_{adj} = P_{raw} \frac{m}{i}$$

where P_{adj} is the adjusted P-value, P_{raw} is the raw P-value, m is the number of tests performed (the number of gene sets in WikiPathways) and i is the rank of the raw P-value among all P-values (where 1 is the smallest P-value, and m is the highest). The odds ratio is how much more likely a gene is to appear in your input list if it belongs to a specific term, compared to a random gene from the background. Genes is the list of genes in the overlap between the genes associated to the term and the genes in the input set.

2.8 Implementation details

The original ScReNI pipeline is written in R. For the purpose of the following experiments, the original code was rewritten into Python, which was preferred for its versatility and usefulness in the development process. The resulting pipeline was called pyScReNI, and produces results with less than 1% difference from the original due to the different seed strategy for random functions and the use of different implementations of library functions. The pipeline was run on the Delft AI Cluster (also known as DAIC) by using 8 CPUs per task and 32 GB of memory for each run. The random forest inferrer takes approximately 5 hours to complete computing (for gwScReNI and type-specific ScReNI) while the analysis takes approximately 20 minutes to compute.

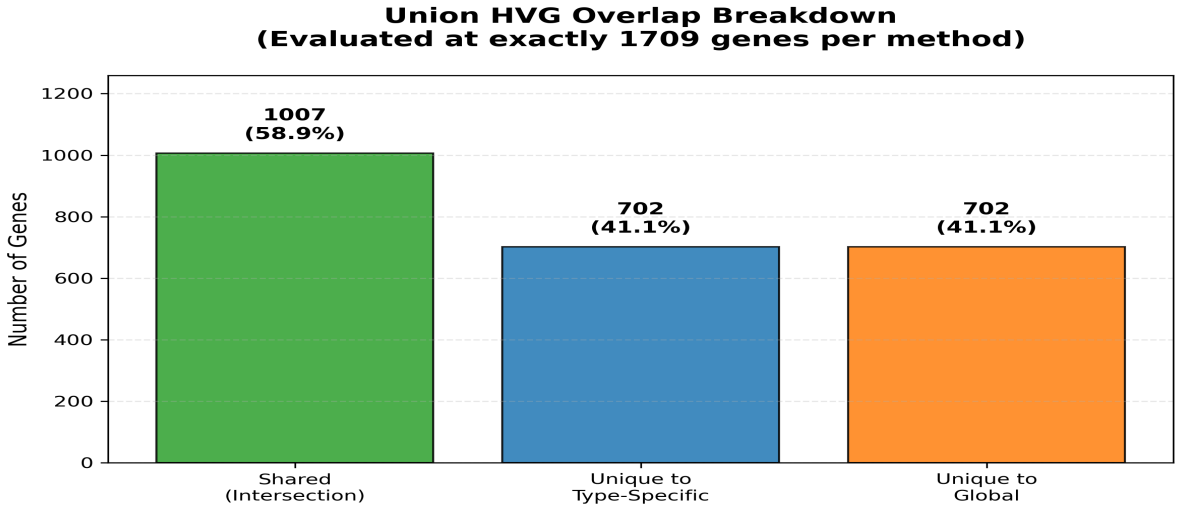


Figure 1: Overlap between global and type-specific HVGs. The green bar represents the genes that the two approaches share; the blue bar represents the genes that only appear in the type-specific HVGs; the orange bar represents the genes that are only taken by the global approach.

3 Results

The ScReNI pipeline was run as close as possible to the original implementation so that the only differences between the two approaches, type-specific and global HVG selection, are the only differentiating factors in the output discrepancies of the two methods. The output was then analyzed in terms of precision/recall, modularity, edge weight and gene set enrichment.

3.1 HVG overlap

In order for the difference in HVG selections to be made apparent, the overlap of genes between type-specific selection and global selection have been plotted. This difference in selection is important to visualize, as it is the only difference that determines the discrepancies in the output of type-specific wScReNI and gwScReNI. The first plot (Figure 1) shows that the two approaches still share a majority of genes (1007 genes - 58.9%) as well as the number of genes (702 - 41.1%) unique to each approach. Figure 2 shows the difference in each type, with RPC1 and RPC3 having 77% and 73.2% in common respectively, with the rest being unique to the type-specific selection, while the MG and RPC2 types have 52.6% and 47.2% in common respectively. This shows that the selection methods truly choose different genes meaning the output of these methods will also be different enough to be compared.

3.2 ChIP-seq edges present in the HVGs of analyzed cell-types

The filtered ground truth contains edges for genes that may not be in the same HVGs, and if so type-specific wScReNI has no way of finding them. This is why the amount of edges present between the genes of each cell-type will give a visualization of the ground-truth data and how different cell-types are represented in the filtered ChIP-seq. Figure 3 illustrates the number of ground-truth edges present between the genes in the cell-specific HVGs of the analyzed cell-types. MG and RPC2 have significantly fewer available true edges present in their HVG-space with MG at 1988 and RPC2 with 662. RPC1 and RPC3 on the other hand have more than double, if not triple that number, with RPC2 having 4868 and RPC1 with 6599 ground-truth edges present. It can therefore be said that, given that the algorithm has fewer "correct choices" for MG and RPC2, the metrics of evaluation for those cell-types will be lower than the ones for RPC1 and RPC3.

3.3 Precision/Recall at different Thresholds

The evaluation of the GRNs starts with precision and recall against ChIP-seq ground-truth. Figure 4 presents the way precision and recall evolve when taking different percentage thresholds of the GRNs. The graph peaks in precision at 1% threshold for type-specific wScReNI and begins to drop with a

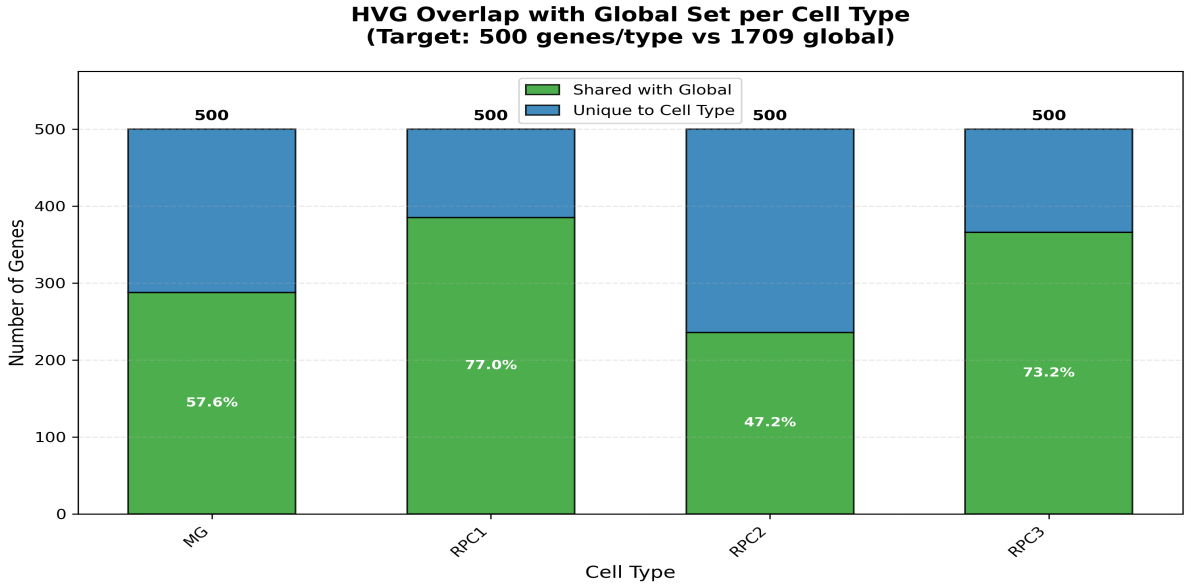


Figure 2: Percentages of genes in each cell-type that the global approach also contains. The green bars represent the overlap in each cell-type of the genes in the type-specific HVGs. The blue bars represent the genes in each cell-type that are unique to the type-specific approach.

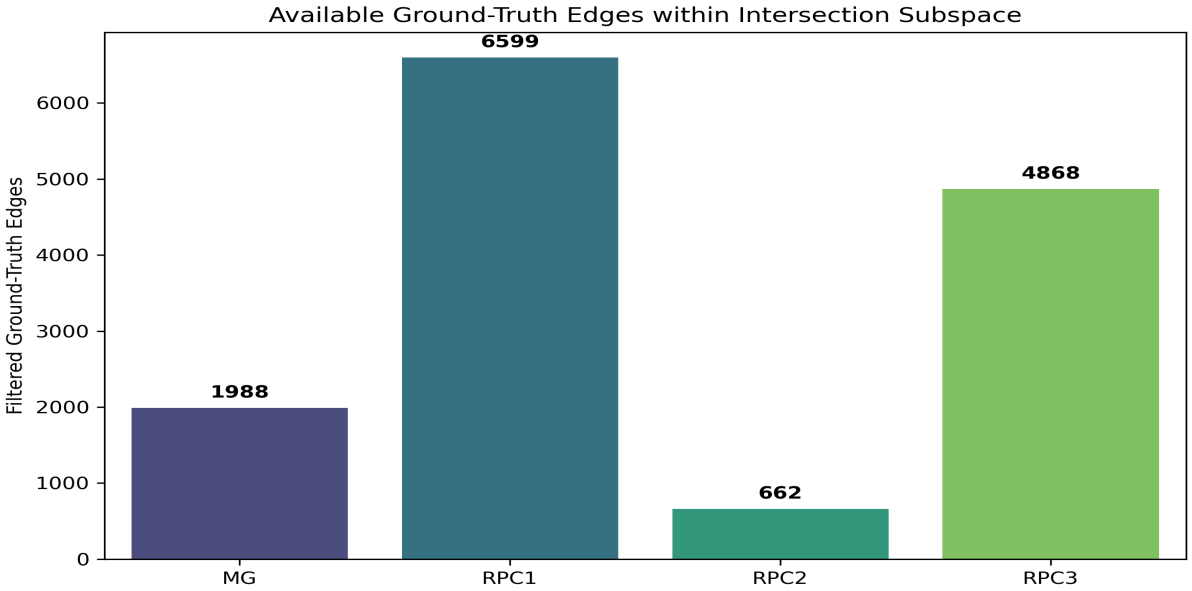


Figure 3: Number of ground-truth edges that are present in the type-specific HVGs of each cell-type. The purple graph shows the number of edges present in ChIP-seq among the 500 HVGs of the MG cell-type; dark blue shows the edges in the HVGs of RPC1; teal shows the edges in RPC2 and the green bar shows the edges in the HVGs of RPC3.

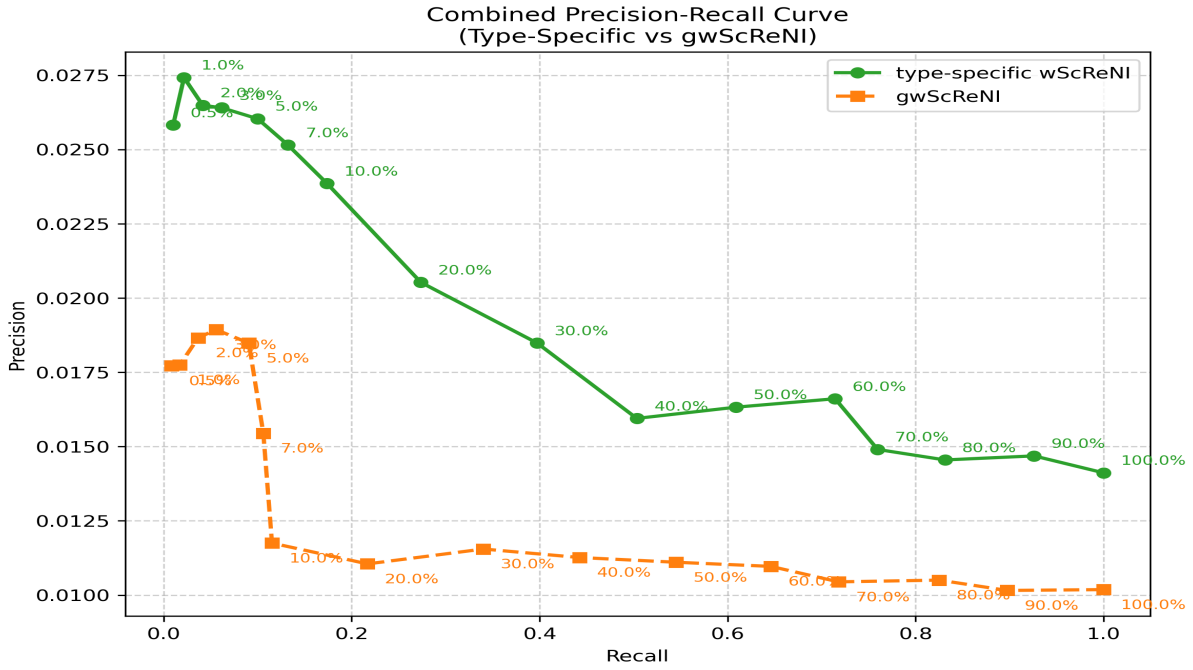


Figure 4: Precision/Recall curve of type-specific wScReNI and gwScReNI at different thresholds. The thresholds are taken at certain percentages of the gene-space of the methods (500x500 for type-specific wScReNI and 1709x1709 for gwScReNI). The precision and recall are evaluated as specified in Methodology against the filtered ChIP-seq ground-truth. The thresholds taken are 0.5%, 1%, 2%, 3%, 5%, 7%, 10%, and from then on until 100% in increments of 10%

minor spike at 60% threshold until 100% of the edges predicted are taken (which is all edges in the gene subspace) where it reaches a recall of 1.0. The graph also shows that gwScReNI is being outperformed by type-specific wScReNI in both precision and recall with gwScReNI peaking in precision at 3% threshold and then stabilizing after 10% threshold. The optimum threshold of the methods is their individual graph spike (1% for type-specific wScReNI and 3% for gwScReNI).

3.4 Cell-type specific precision/recall curves

The reason for the results in Figure 4 can be attributed to a better evaluation of the specific cell-types by type-specific wScReNI as shown in Figure 5. This figure shows the evolution of precision and recall over multiple thresholds for each individual cell-type: RPC1 and RPC3 have almost double the amount of precision in type-specific wScReNI compared gwScReNI, while MG and RPC1 have a slightly lower precision than their gwScReNI counterparts. This can be attributed to their decreased number of ground-truth edges (as evidenced in Figure 3) leading to an overestimation of the algorithm regarding the weights it attributes to false positive edges. All cell-types in gwScReNI have a similar graph as is expected, given that the global approach does not take cell-type into account when calculating the importance of gene to gene edges in the random forest. This inevitably leads to similar GRNs for cells of different cell-types with similar gene expressions. The different shapes of the cell-types in type-specific wScReNI can also be attributed to the fact that the random forest only takes the 500 HVGs of that cell-type into account when calculating the importance of edges.

3.5 Weight distribution of edges

Another reason for the increased precision and recall is the difference in weight distribution shown in Figure 6. As can be seen, type-specific wScReNI dominates gwScReNI in weight values calculated until the 1% threshold. For weights close to 0, gwScReNI dominates type-specific and because of the log scale of the density it means gwScReNI has a significantly higher density of close to 0 weighted edges. This is because of the global nature of the selection, the GRNs have a higher gene-space (1709x1709) which naturally leads to a more "flat" distribution of edges. Although a higher weight does not mean a necessarily greater confidence of the algorithm in that edge, the algorithm takes those edges earlier in

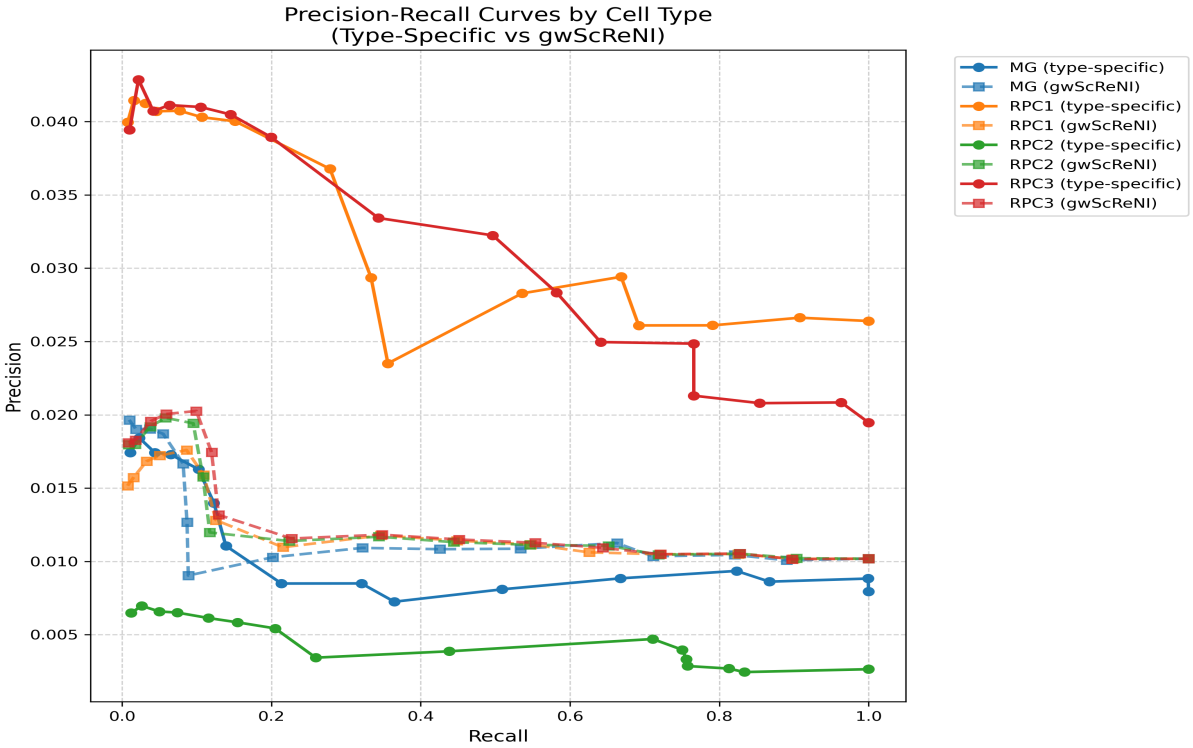


Figure 5: Precision/Recall curves of the cell-types analyzed at different thresholds for type-specific wScReNI (solid lines) and gwScReNI (dotted lines). The cell-types have consistent colors between the two methods. The thresholds taken are the same as in Figure 4.

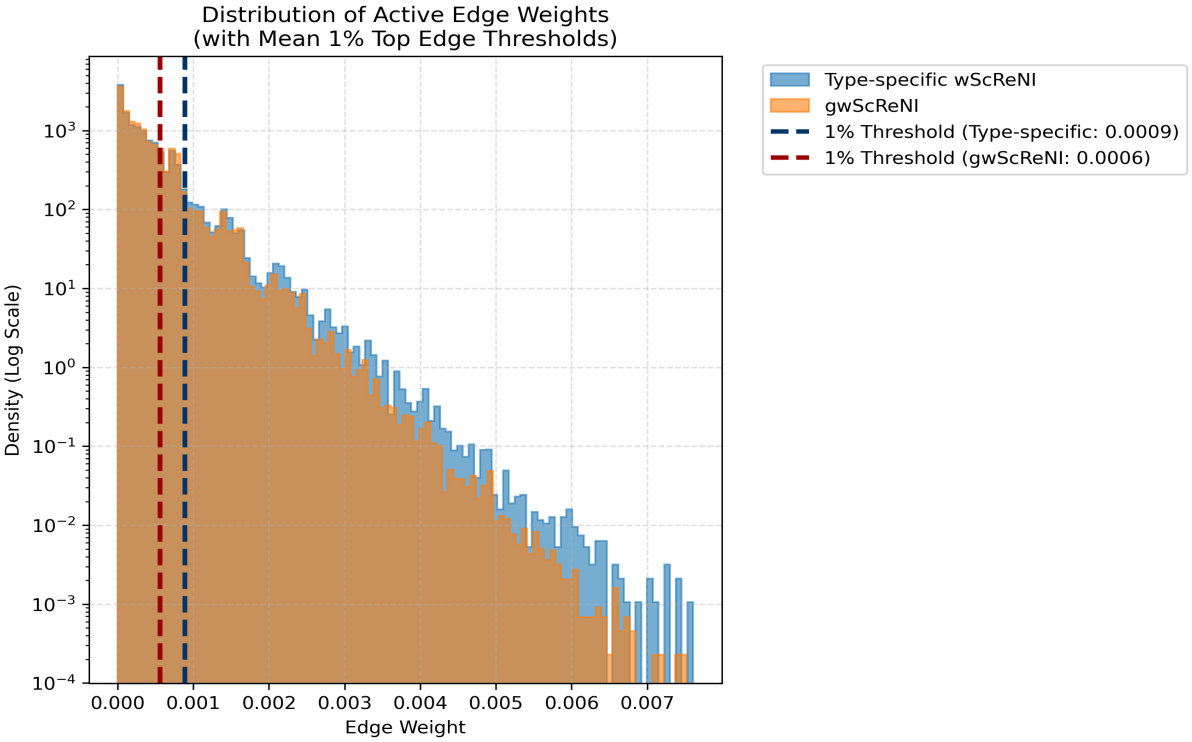


Figure 6: Weight distribution histogram of edges in gwScReNI and type-specific wScReNI. The density (vertical) axis is in log scale. The 1% threshold is also plotted and the exact weight value that constitutes that threshold is specified next to it in the legend.

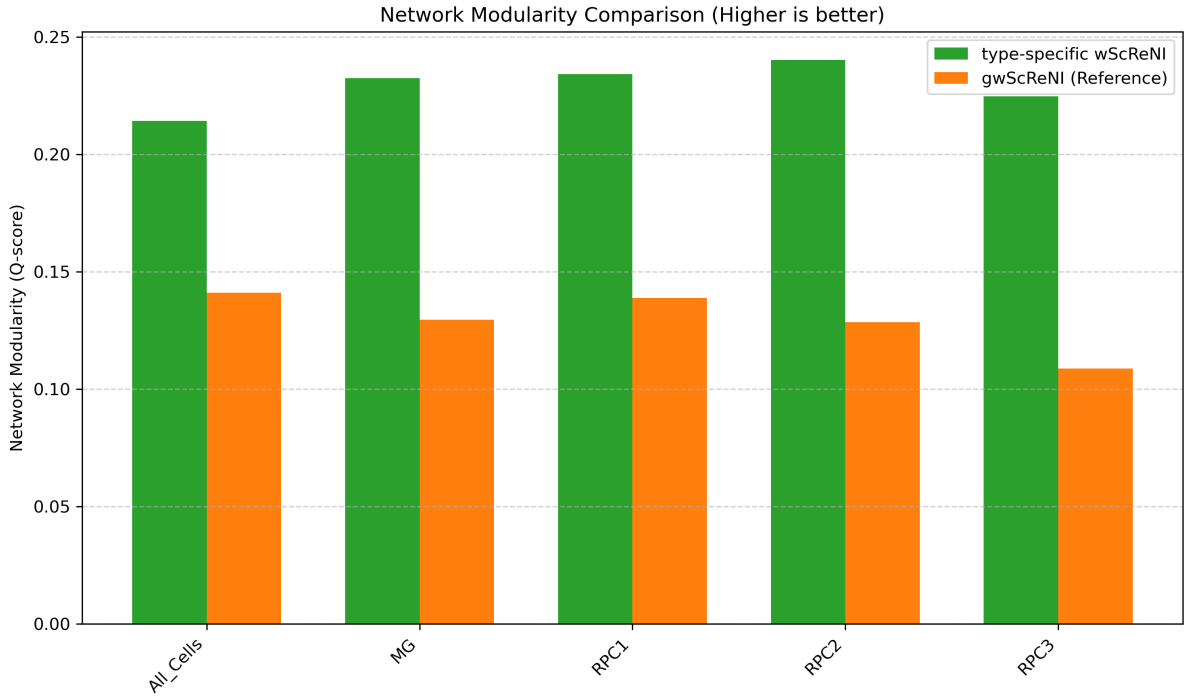


Figure 7: Modularity metric of type-specific wScReNI compared to gwScReNI. A higher Q means a more strongly modular network. The calculation of this metric is detailed in the Methodology section. All_Cells represents the modularity of the mean GRN of all cells analyzed, while the Q of the individual cell-types is calculated on the mean of GRNs of that cell-type. Only the top 1% of edges (by decreasing weight) are considered for the mean GRNs.

the thresholds and spends a longer time with those edges, rather than the ones for which the random forest inferred an importance close to 0.

3.6 Modularity of the GRNs

For the biological evaluation the Modularity (Q) was calculated for the GRNs (Figure 7). A higher Q means a more modular network, which is in line with the biological truth of GRNs[11] as described in the Methodology section. Type-specific wScReNI consistently achieves a higher Q-score than gwScReNI, showing that the GRNs are structurally different when applying the type-specific approach rather than the global one. If gene set enrichment validates these results, it would mean that type-specific wScReNI generates GRNs with more biological significance of the pathways that occur inside of cells.

3.7 Gene set enrichment

For the purpose of validating the modules found, gene set enrichment was run as mentioned in the methodology. The findings (Table 1 in the appendix) show that the modules produced by type-specific wScReNI have almost no biological significance, relying on 1-3 genes that overlap with the pathways and having high Adjusted P-values. This leads to a noisy output, with pathways such as Alzheimer's or Parkinson's Disease being found, which have little significance for the mouse retinal development data the algorithm was run on. Among the statistically (meaning Adjusted P-value < 0.05) and biologically significant findings in type-specific wScReNI there is: *Mechanisms Associated With Pluripotency WP1763* in Module 7 of RPC1, which line up with the biological identity of progenitor (RPC) cell states[9], with an Adjusted P-value of 0.039, but only two overlap genes (PSEN1 and EZH2); and *Mapk Signaling Pathway WP493* in Module 2 of MG, which is involved in cellular responses to stress and survival[24], with an Adjusted P-value of 0.37, but only four overlap genes (CASP7, NLK, MAPT and FOS). The low number of overlap genes, the high adjusted P-values and the finding of seemingly statistically significant unrelated disease pathways (Parkinson in Module 2 of MG and Alzheimer's in Module 7 of RPC1) shows that the biological significance is fragile and can be misleading.

On the other hand, gwScReNI finds pathways with more statistical significance, backed by a wider gene overlap (Table 2 in the appendix). *Cytoplasmic Ribosomal Proteins WP163* in Module 0 of RPC1, is a signature of proliferation in progenitor cell-types[17], with an Adjusted P-value of 8.33×10^{-12} and 23 overlapping genes (RPS and RPL). This pathway is also found with significance in Module 1 of MG, Module 1 of RPC2 and Module 1 of RPC3, having an overlap of 18, 24 and 16 genes respectively. *Dopaminergic Neurogenesis WP1498* and *Neural Crest Differentiation WP2074* are also found in Module 3 of RPC3, with an overlap of 7 and 8 genes respectively, indicating the progenitor cells at the end of the cell cycle becoming photoreceptors or amacrine cells[4], identifying a different module with a different biological function. This does imply a deeper biological significance, but no other important biological functions have been found, which can indicate a bias by gwScReNI towards these genes.

3.8 GRN plots

In order to find if gwScReNI truly is biased towards RPS genes, the GRNs of gwScReNI and type-specific wScReNI were plotted in Figure 9 for MG and RPC1 and in Figure 10 for RPC2 and RPC3. The GRNs of gwScReNI for all four cell types have as the center hub regulator the gene RPS16 (Ribosomal Protein S16) with RPS24 appearing as a regulator as well in all cell-types but RPC2. These are known false regulators, as RPS16 and RPS24 are so-called "housekeeping" genes[10] whose protein products are necessary for the cell's survival and are therefore expressed at a much higher rate than true regulators. This leads to GRN inferring algorithms mistaking these genes as important hubs and missing real regulators[16]. Meanwhile, type-specific wScReNI finds Nr2c2cap as the main regulator for MG[22]; Fgfr1op[2] and Sat1[19] as regulators for RPC1; Zfp821[14] and Lyar[20] as regulators for RPC2; Nfib[6] as a regulator for RPC3. All these are known regulators for their respective cell-types and show that type-specific wScReNI has significant biological importance in its inferred GRNs

4 Responsible Research

The work on single-cell GRN algorithms is for the advancement of the field of bioinformatics which inherently raises ethical considerations. On the one hand, these algorithms can be used to find cures for diseases and thereby saving and improving countless lives, which is the purpose of this paper. On the other hand, this knowledge can be used for less noble purposes that would be detrimental to society. The authors condemn these actions and advocate strictly for the ethical use of this technology. The algorithm also has a carbon footprint that contributes to the ongoing climate change[1].

In this paper, generative AI, namely Gemini, was used only for coding, summarizing information and finding references. While coding, it was used to implement methods and functions for plotting, fixing bugs and adding comments. It had constant supervision by the authors, and mistakes (hallucinations) that it introduced were found and fixed. It was also used to summarize data that was too vast for human analysis as well as formatting and correcting the grammar on written parts of this project and plots. When it generated information that the authors were not familiar with, it was asked for academic sources regarding that information and the authors checked those references. Experiments, coding and academic decisions were all made by the authors with the help of the constant support of the supervisors. The responsibility of correctness, originality and integrity lies solely with the authors of this paper.

The reproducibility of the experiments undertaken in this paper was ensured by having set seeds for inherently random functions. There is a `README.md` file that will guide users on how to run the project and where to find the data necessary for the analysis. Most of the code tied to ScReNI was developed by Xu et al. in their paper[21] in R and the data necessary for the comparison (Mouse retinal scRNA-seq and scATAC-seq; ChIP-seq ATLAS) is available in their repository `https://github.com/Xuxl2020/ScReNI`. The authors of this paper rewrote the code into Python and added the modifications necessary for the experiments that were conducted. The experiments were run (as mentioned in Methodology) on the Delft AI Cluster and contain jobs for that purpose in the *slurm* directory of the project. The entire project alongside all necessary information on reproducing the results is available at `https://github.com/MohneaGosu/RP-project`.

5 Discussion

The main challenge to this algorithm is the lack of a ground truth. ChIP-seq is assumed to be one, but as Datta et al. have shown[8] it can have biases and a high count of false positives. Given this, the results of precision and recall should also not be taken as absolute. Nevertheless, ChIP-seq is still widely used

and considered the industry standard[15] and Xu et al. use it in the original ScReNI[21] paper as well. Therefore, an increase in precision/recall is still important to show a better performance of an approach over another. Recall as seen in the graph is also heavily dependent on the threshold taken, as simply predicting a bigger GRN will naturally increase the recall at the expense of the precision, with 100% threshold having a recall value of 1.0 while precision drops to the ratio of true ChIP-seq edges to the total possible edges in the gene-space.

On the other hand, the modularity is less biased, but does not give much information about the validity of the network. The fact that the enrichment of the modules of type-specific wScReNI came out as inconclusive, giving only statistically fragile and biologically mixed data, shows that simply having a high Q-score is not enough to guarantee good results. The enrichment of gwScReNI data being more significant also shows the shortcomings of a purely precision-recall oriented evaluation. For meaningful enrichment to be conducted on the modules, a larger set of genes has to be used, but given that the ScReNI pipeline has at best a $O(N^2)$ complexity (given that the output matrix of the random forest is a $N \times N$ matrix) a large number of HVGs is computationally expensive and leads to more and more noise.

6 Conclusion

In this work, we compared two different HVG selection approaches for the ScReNI pipeline: a global approach and a type-specific approach. The type-specific approach performed significantly better in precision and recall, weight distribution, modularity (Q-score) and regulator prediction, while the global approach had better results for gene set enrichment of the modules, managing to generate modules with different biological functions. Given these results we concluded that type-specific approach performed better at inferring the single-cell GRNs by leveraging the data that was most significant to the single cells - the cell-type specific genes.

Future additions in this area should investigate the cause of the discrepancy in available ground-truth edges present in the gene spaces of the cell types and a more reliable ground truth variant than ChIP-seq. Additionally, it would be of interest to find even more computationally optimal ways of inferring single-cell GRNs as well as mitigating the noise introduced by highly expressed "housekeeping" genes[10].

A Appendix

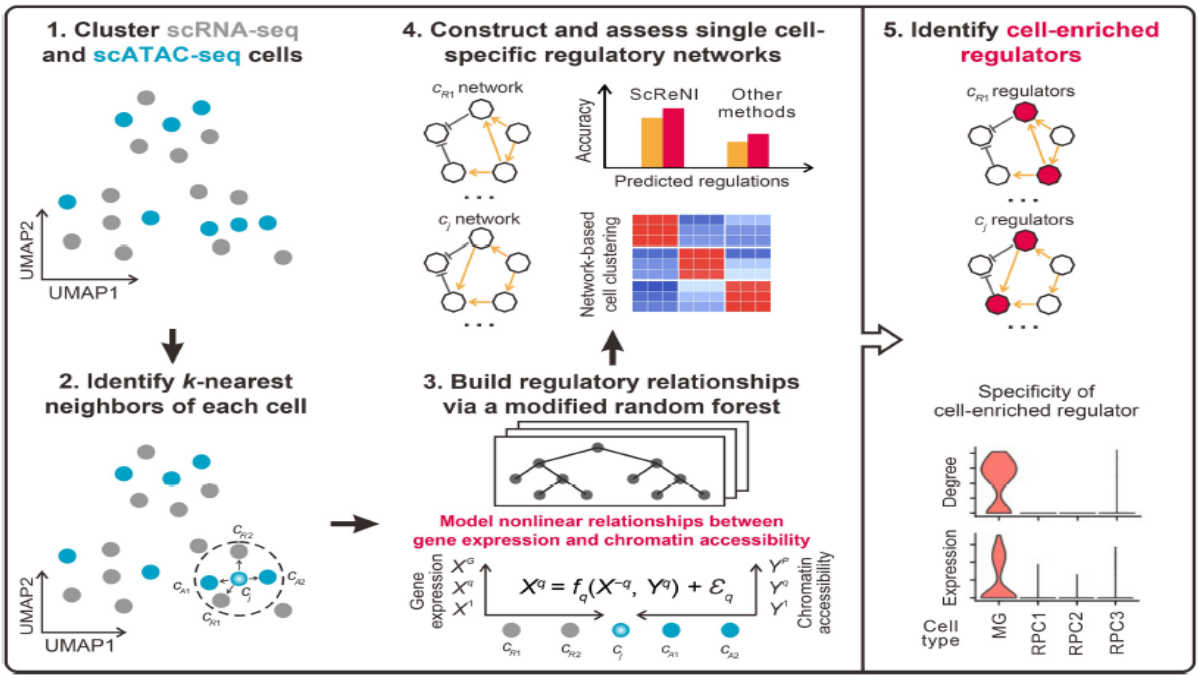


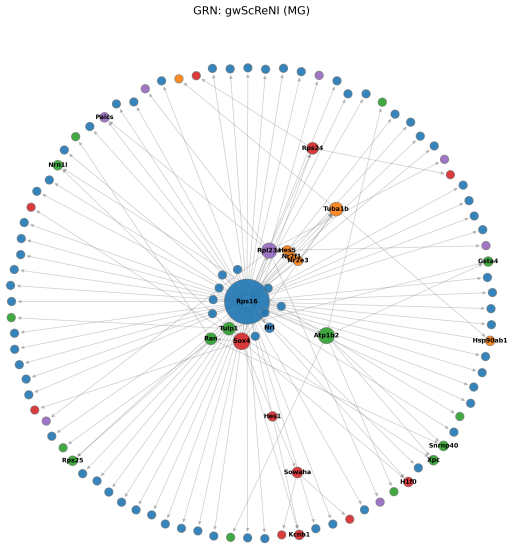
Figure 8: ScReNI[21] pipeline: **1.** The cells are clustered and the RNA data is paired with similar ATAC data; **2.** k-nearest neighbors are computed using UMAP; **3.** The random forest learns the function f_q based on the neighbors of gene q, here denoted as X^{-q}, the peaks associated with gene q, here denoted as Y^q and a random noise ϵ_q; **4.** The GRNs are constructed from the weights inferred by the random forest; **5.** Plot and analyze the GRNs

Table 1: Pathway Enrichments of type-specific wScReNI modules. The data was truncated to only what is used in the Gene set enrichment subsection of Results. The full data is available on the repositories mentioned in the Responsible Research section

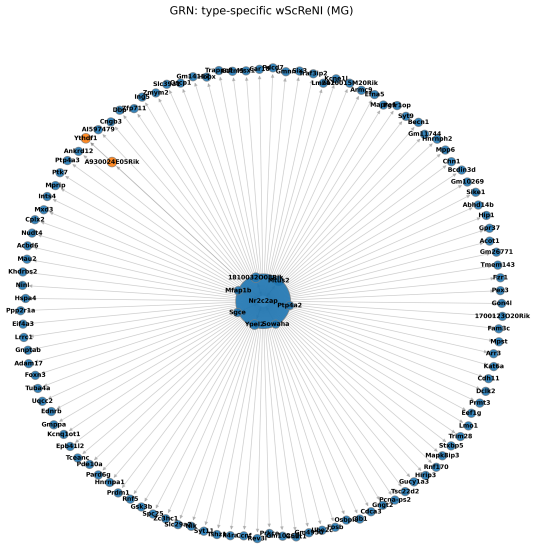
Module (Cell-Type)	**Term**	**Adjusted P-value**	**Odds Ratio**	**Genes**
Module 2 (MG)	Parkinson 39 S Disease WP3638	0.022530	43.143791	GPR37; CASP7; SYT11
Module 2 (MG)	Mapk Signaling Pathway WP493	0.036852	11.400000	CASP7; NLK; MAPT; FOS
Module 7 (RPC1)	Alzheimer 39 S Disease WP2075	0.039521	60.428571	PSEN1
Module 7 (RPC1)	Mechanisms Associated With Pluripotency WP1763	0.039521	23.243902	PSEN1; EZH2

Table 2: Pathway Enrichments of gwScReNI modules. The data was truncated to only what is used in the Gene set enrichment subsection of Results. The full data is available on the repositories mentioned in the Responsible Research section

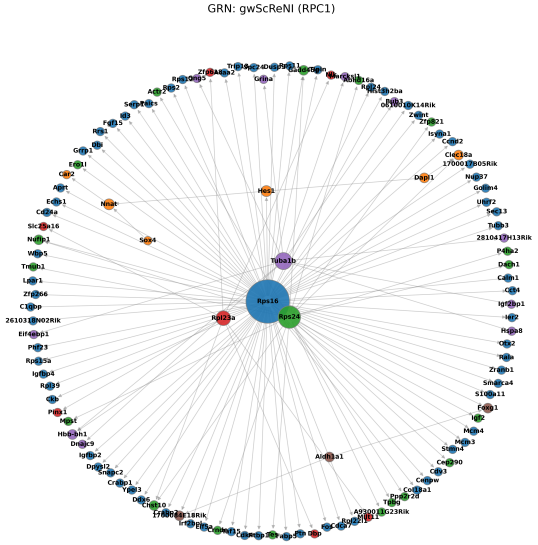
Module (Cell-Type)	**Term**	**Adjusted P-value**	**Odds Ratio**	**Genes**
Module 0 (RPC1)	Cytoplasmic Ribosomal Proteins WP163	8.328130e-12	17.200633	RPS12; RPS17; RPLP1; RPL34; RPL39; RPL4; RPL3; RPS11; RPL10A; RPS19; RPS3A1; RPS4X; RPL32; RPS25; RPS8; RPL21; RPS15A; RPL24; RPL12; RPL10; RPS27A; RPS23; RPL7
Module 1 (MG)	Cytoplasmic Ribosomal Proteins WP163	0.013495	4.391584	RPL24; RPL12; RPS23; RPL4; RPS4X; RPL32; RPS12; RPS17; RPS8; RPLP1; RPL6; RPS11; RPL10A; RPL39
Module 1 (RPC2)	Cytoplasmic Ribosomal Proteins WP163	7.121086e-12	17.707457	RPS12; RPS17; RPLP1; RPL34; RPL26; RPL39; RPL4; RPL6; RPS11; RPL10A; RPS19; RPS4X; RPL32; RPS25; RPS8; RPL21; RPS15A; RPL24; RPL12; RPL10; RPS27A; RPS23; RPS2; RPL7
Module 1 (RPC3)	Cytoplasmic Ribosomal Proteins WP163	0.000182	6.275618	RPL24; RPS19; RPL21; RPL12; RPS23; RPL4; RPS27A; RPS4X; RPS2; RPL32; RPS12; RPS8; RPL6; RPLP1; RPL34; RPS11
Module 3 (RPC3)	Dopaminergic Neurogenesis WP1498	0.003688	14.761352	PITX3; NEUROD1; ASCL1; SOX2; MSX1; NEUROG2; OTX2
Module 3 (RPC3)	Neural Crest Differentiation WP2074	0.007757	7.822378	DLL3; DLL1; ID1; ASCL1; ISL1; ZIC1; SOX5; TFAP2B



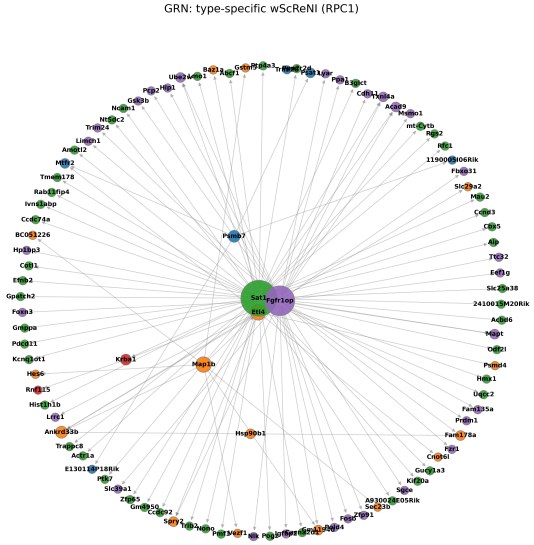
(a) gwScReNI mean GRN of MG cell-type



(b) type-specific mean GRN of MG cell-type

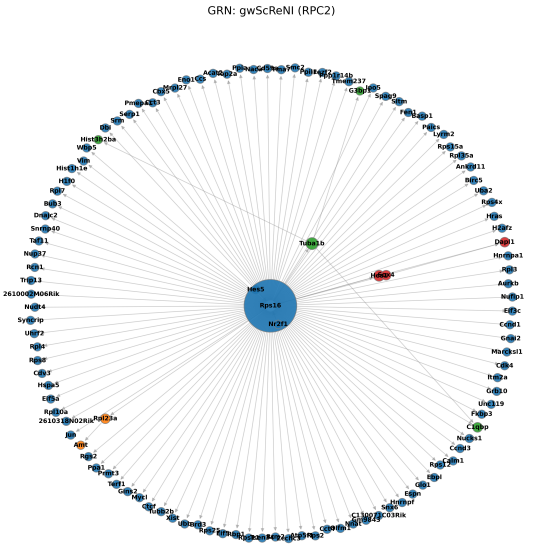


(c) gwScReNI mean GRN of RPC1 cell-type

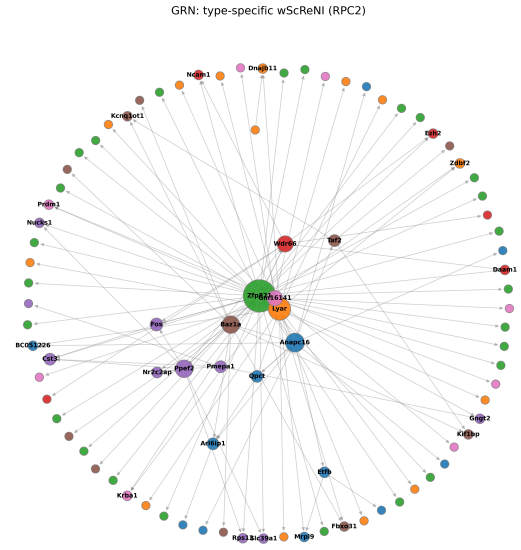


(d) type-specific mean GRN of RPC1 cell-type

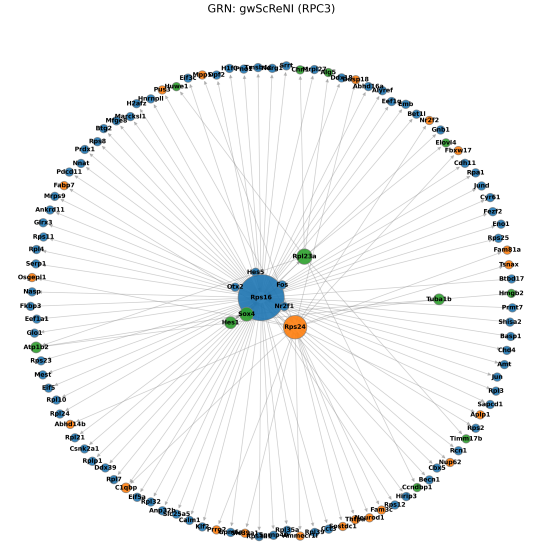
Figure 9: Part 1: Plots of mean GRN for MG and RPC1 cell-types for gwScReNI and type-specific ScReNI. The GRNs are calculated at 0.005% threshold of possible type-specific edges. Colors represent the modules present in the network. The size of the nodes is proportional to their degree.



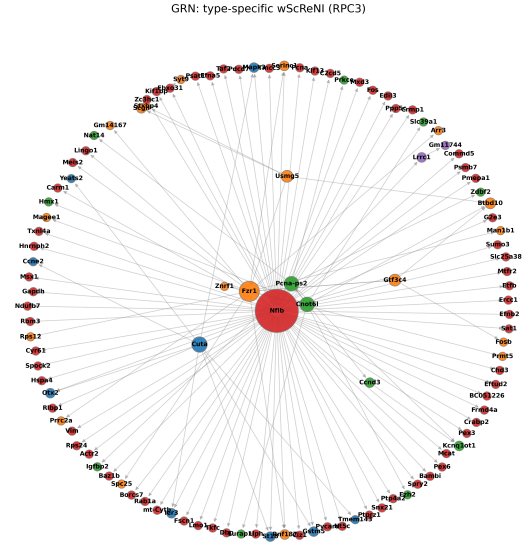
(a) gwScReNI mean GRN of RPC2 cell-type



(b) type-specific mean GRN of RPC2 cell-type



(c) gwScReNI mean GRN of RPC3 cell-type



(d) type-specific mean GRN of RPC3 cell-type

Figure 10: Part 2: Plots of mean GRN for RPC2 and RPC3 cell-types for gwScReNI and type-specific ScReNI. The GRNs are calculated at 0.005% threshold of possible type-specific edges. Colors represent the modules present in the network. The size of the nodes is proportional to their degree.

References

[1] Kashif Abbass, Muhammad Zeeshan Qasim, Huaming Song, Muntasir Murshed, Haider Mahmood, and Ijaz Younis. A review of the global climate change impacts, adaptation, and sustainable mitigation measures. *Environmental science and pollution research*, 29(28):42539–42559, 2022.

[2] Claire Acquaviva, Véronique Chevrier, Jean-Paul Chauvin, Gaëlle Fournier, Daniel Birnbaum, and Olivier Rosnet. The centrosomal fop protein is required for cell cycle progression and survival. *Cell cycle*, 8(8):1217–1227, 2009.

[3] Ayushi Agrawal, Hasan Balcı, Kristina Hanspers, Susan L Coort, Marvin Martens, Denise N Slenter, Friederike Ehrhart, Daniela Digles, Andra Waagmeester, Isabel Wassink, et al. Wikipathways 2024: next generation pathway database. *Nucleic acids research*, 52(D1):D679–D689, 2024.

[4] Revathi Balasubramanian and Lin Gan. Development of retinal amacrine cells and their dendritic stratification. *Current ophthalmology reports*, 2(3):100–106, 2014.

[5] Zeinab Breijyeh and Rafik Karaman. Comprehensive review on alzheimerâs disease: causes and treatment. *Molecules*, 25(24):5789, 2020.

[6] Brian S Clark, Genevieve L Stein-OâBrien, Fion Shiau, Gabrielle H Cannon, Emily Davis-Marcisak, Thomas Sherman, Clayton P Santiago, Thanh V Hoang, Fatemeh Rajaii, Rebecca E James-Esposito, et al. Single-cell rna-seq analysis of retinal development identifies nfi factors as regulating mitotic exit and late-born cell specification. *Neuron*, 102(6):1111–1126, 2019.

[7] Hao Dai, Lin Li, Tao Zeng, and Luonan Chen. Cell-specific network constructed by single-cell rna sequencing data. *Nucleic acids research*, 47(11):e62–e62, 2019.

[8] Vishaka Datta, Sridhar Hannenhalli, and Rahul Siddharthan. Chipulate: A comprehensive chip-seq simulation pipeline. *PLoS computational biology*, 15(3):e1006921, 2019.

[9] Emily Davis, Abdullah Khan, and Issam Aldiri. Ezh2 control of bivalent genes fine-tunes developmental competence during retinogenesis. *Investigative Ophthalmology & Visual Science*, 67(6):18–18, 2026.

[10] Eli Eisenberg and Erez Y Levanon. Human housekeeping genes, revisited. *TRENDS in Genetics*, 29(10):569–574, 2013.

[11] Claus Kadelka, Matthew Wheeler, Alan Veliz-Cuba, David Murrugarra, and Reinhard Laubenbacher. Modularity of biological systems: a link between structure and function. *Journal of the Royal Society Interface*, 20(207):20230505, 2023.

[12] Philipp Keyl, Philip Bischoff, Gabriel Dernbach, Michael Bockmayr, Rebecca Fritz, David Horst, Nils Blüthgen, Grégoire Montavon, Klaus-Robert Müller, and Frederick Klauschen. Single-cell gene regulatory network prediction by explainable ai. *Nucleic Acids Research*, 51(4):e20–e20, 2023.

[13] Marieke Lydia Kuijjer, Matthew George Tung, GuoCheng Yuan, John Quackenbush, and Kimberly Glass. Estimating sample-specific regulatory networks. *Iscience*, 14:226–240, 2019.

[14] Samuel A Lambert, Arttu Jolma, Laura F Campitelli, Pratyush K Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R Hughes, and Matthew T Weirauch. The human transcription factors. *Cell*, 172(4):650–665, 2018.

[15] Stephen G Landt, Georgi K Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E Bernstein, Peter Bickel, James B Brown, Philip Cayting, et al. Chip-seq guidelines and practices of the encode and modencode consortia. *Genome research*, 22(9):1813, 2012.

[16] Wei Vivian Li and Yanzeng Li. sclink: inferring sparse gene co-expression networks from single-cell expression data. *Genomics, proteomics & bioinformatics*, 19(3):475–492, 2021.

[17] Soyeon Lim, You-Joung Kim, Sooyeon Park, Ji-heon Choi, Young Hoon Sung, Katsuhiko Nishimori, Zbynek Kozmik, Han-Woong Lee, and Jin Woo Kim. mtorc1-induced retinal progenitor cell overproliferation leads to accelerated mitotic aging and degeneration of descendent müller glia. *Elife*, 10:e70079, 2021.

[18] Malte D Luecken and Fabian J Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):MSB188746, 2019.

[19] Anthony E Pegg. Spermidine/spermine-n 1-acetyltransferase: a key metabolic regulator. *American Journal of Physiology-Endocrinology and Metabolism*, 294(6):E995–E1010, 2008.

[20] Lishan Su, R Jane Hershberger, and Irving L Weissman. Lyar, a novel nucleolar protein with zinc finger dna-binding motifs, is involved in cell growth regulation. *Genes & development*, 7(5):735–748, 1993.

[21] Xueli Xu, Yanran Liang, Miaoxiu Tang, Jiongliang Wang, Xi Wang, Yixue Li, and Jie Wang. Screni: single-cell regulatory network inference through integrating scrna-seq and scatac-seq data. *Genomics, Proteomics & Bioinformatics*, 23(4):qzaf060, 2025.

[22] Yue Yang, Xin Wang, Tiefei Dong, Eungseok Kim, Wen-Jye Lin, and Chawnshang Chang. Identification of a novel testicular orphan receptor-4 (tr4)-associated protein as repressor for the selective suppression of tr4-mediated transactivation. *Journal of Biological Chemistry*, 278(9):7709–7717, 2003.

[23] Qiuyue Yuan and Zhana Duren. Inferring gene regulatory networks from single-cell multiome data using atlas-scale external data. *Nature Biotechnology*, 43(2):247–257, 2025.

[24] Samuel Shao-Min Zhang, Hong Li, Ping Huang, Lucy Xi Lou, Xin-Yuan Fu, and Colin J Barnstable. Mapk signaling during müller glial cell development in retina explant cultures. *Journal of ocular biology, diseases, and informatics*, 3(4):129–133, 2010.

[25] Ziqi Zhang, Jongseok Han, Le Song, and Xiuwei Zhang. Cespgrn: inferring cell-specific gene regulatory networks from single cell multi-omics and spatial data. *bioRxiv*, pages 2022–03, 2022.

[26] Zhaonan Zou, Tazro Ohta, and Shinya Oki. Chip-atlas 3.0: a data-mining suite to explore chromosome architecture together with large-scale regulome data. *Nucleic Acids Research*, 52(W1):W45–W53, 2024.