**Abbreviation list:**

IS: Interaction score

NIS: normalized interaction score

scRNA-seq single cell RNA sequencing

TF: transcription factor

**Introduction:**

* Exhaustion
* Nina Hepatitis C paper
* GRN
* Binarized network
* Refer to Bolouri?
* Refer to Herault?

**pySCENIC** **Preparation: 102**

To run pySCENIC an input table with gene expression in each cell is required. From the raw single cell RNA sequencing (scRNA-seq) data only expression information in cells that had a cell type assigned was kept. Genes, that did not meet the following filtering criteria were left out:

**sum of counts for a single gene among all cells >2\*0.01\*2\*0.01\*number of cells**

Some differences in nomenclature between our scRNA-seq data and Boulori paper network were identified and subsequently corrected. It was made sure that the maximum amount of Bolouri paper network nodes was included in the pySCENIC input. Out of 63 nodes, 48 were already added during the gene filtering step and 8 more were added manually afterwards. As a result, a table containing expression data of 10 245 genes among 784 cells was used as pySCENIC input.

pySCENIC provides three files containing information about, importance of single interactions between a transcription factor (TF) and a target, composition of a regulon and regulon activity in studied cells. In order to obtain more robust regulons and interactions a total amount of 50 pySCENIC iterations was performed.

**pySCENIC** **clean up: 173**

Each pySCENIC iteration generates an AUCell output file which consists of a table where activity of TFs is quantified in each cell. These tables were used to create a reference subset of unique TFs. In total, pySCENIC identified different TFs among 50 runs. These TFs were be used in the next step to filter for only TF-TF interactions.

The RcisTarget output file mainly contains information about the regulons, where multiple targets are listed for each TF. Two version of the regulons were kept. The first one, without any filtering, containing all possible genes identified as targets for a particular TF. The second one, had its targets reduced to only TF-TF interactions, which amounted to 6 254 unique interactions. Further analysis and clean up centers on the second version of the regulons.

To reduce the scope of interactions to be analyzed, only interactions that appeared in 40 or more pySCENIC runs (80% of the runs) were kept. Consequently, 641 interactions were left.

For these interactions NIS score was calculated, where for a target gene t with j regulators r all j regulators are considered.

**NIS(ri, t) = IS(ri, t)/SUM((rj, t))**

IS(ri, t) is an average of all the importance scores for this interaction from grnboost2 pySCENIC output files multiplied by the number of pySCENIC runs, in which interaction from ri, to t was identified. NIS allows to give more relevancy to interactions that are recovered more frequently and therefore is an option for weighting them.

**pySCENIC** **interactions confirmation: 373**

Various literature sources were considered to validate pySCENIC interactions. In this project databases Cistrome and Dorothea were chosen. Additionally, ATAC-seq data from patients who recovered from hepatitis C viral infection was used as another source of confirmation.

Cistrome

Out of all TFs presented in Cistrome database only the ones that were classified as “Blood” in the tissue type category were picked from the Cistrome database for further analysis. As the result, a total of 23% of TFs from the pySCENIC subset were selected (80 TFs). For each TF 6000 best peaks according to the ranking in the files were chosen. Each of the peaks was expanded 1 kb in both directions in order to overlap it with human genome. Human genes that would lie in this extended window were classified as confirmed targets of this particular TF. In total 20% of interactions that made it through the frequency filter were confirmed by Cistrome using this approach.

Dorothea: 497

Dorothea was used a second reference database. Out of all overlapping interactions only the ones that had “A” confidence were kept. Thereby, 33 interactions were confirmed.

ATAC-seq: 530

Recovered patients ATAC-seq data contributed to further identification of robust interactions between TFS. Subsequently, 98 interactions were verified.

Summary: 601

Across the 641 most frequent interactions 33% (210 interactions) were confirmed by at least one of mentioned validation sources.

**Binarize activity for TFs:629**

1. Working with full regulons (not only TF-TF interactions, but also including other genes)
2. Keep only regulons that pass frequency filter
3. For these regulons conduct new AUCell
4. Heatmap: new\_aucell\_not\_norm.pdf
5. Random Forest to pick 100 TFs that explain the most our cell types
6. Binarize AUCell score for each cell (depending on the cell type) in those 100 TFs. Percentage of cells that are active in each cell type.
7. Heatmap: aucell\_scores\_norm\_top100\_cell\_type\_binarized\_percentage
8. Assign cell type to each TF

**Network of 100 RF TFs**

1. 90% filter on interactions
2. Network: 82\_RF\_TFs.png

**Overlap our network with Bolouri paper Network**

1. 9 overlapping nodes
2. Network: 41\_paper\_overlap\_TFs.png

**Fusion of our network and Bolouri paper network using common nodes**

1. Adapt the nomenclature and format Bolouri paper interactions
2. Fuse networks
3. Network: 66\_bolouri\_RF\_fusion.png

**Reduce the network to relevant nodes**

1. Explain why
2. Network: 66\_bolouri\_RF\_fusion\_curated
3. Network: 66\_bolouri\_RF\_fusion\_curated\_2

**Further reduce the network to only**

1. Explain why
2. Network: 13\_Bonesis\_toy

**Discussion Points**

* Cistrome peaks and extensions
* RF TFs