**Abbreviation list:**

IS: Interaction score

NIS: normalized interaction score

RF: Random Forest

scRNA-seq single cell RNA sequencing

TF: transcription factor

**Introduction:**

* Exhaustion
* Nina Hepatitis C paper
* GRN
* Binarized network
* Refer to Bolouri?
* Refer to Herault?
* How does pySCENIC work?

**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. There are three different cell types: memory, transitional and exhausted. 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 353 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 among 251 TFs 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, 2% of most frequent interactions were confirmed. Important to note, that none of these interactions were identified via other two validation methods that were used.

ATAC-seq: 530

Recovered patients ATAC-seq data contributed to further identification of robust interactions between TFS. Subsequently, 15% of interactions, that passed the frequency filter, were verified.

Summary: 601

Across the 641 most frequent interactions from 50 pySCENIC runs 33% (210 interactions) were confirmed by at least one of mentioned validation sources.

**Binarize activity for TFs:629**

The following part centers on binarizing the activity of the TFs. Simultaneously the goal is to reduce TFs amount (at this stage 251 TFS) in order to be able to visualize their interactions in a network.

To quantify activity of TFs in the cells AUCell was used. As regulons for each of the TFs its full list of genes (not only restricting TF-TF interactions), that was saved during pySCENIC output processing, was given as AUCell input. Small regulons (< 15 targets) were removed, leaving 222 regulons. The assumption was to get a bimodal distribution and define a cutoff value, that would separate cells where a TF is active and inactive. However, both visual perception and bimodality tests showed that AUCell scores distribution do not meet this expectation **(Figure: get one of the AUCell histograms)**.

Even though there was no bimodality observed, after normalizing the AUCell scores, and activity difference among cell types within a TF could be observed **(Figure: new\_aucell\_norm)**.

To further reduce the number of TFs in consideration Random Forest (RF) was conducted. In this RF the goal was to discover how well do the AUCell scores of the TFs explain the belonging of the cells to one or another cell type. Therefore, each TF becomes and explanatory variable for the cell type (exploratory variable). RF can also deliver an importance score for each explanatory variable, which was the main point of interest in this approach. On the grounds of this, the total TF amount for further analysis was reduced to 100 highest ranked by RF.

For these 100 TFs an average of activity (AUCell scores) within each cell type was calculated **(Sup. Figure: aucell\_scores\_norm\_top100\_cell\_type\_average)**. The distributions of averages were bimodal, as it can be confirmed both visually and by bimodal tests. **(Sup. Figure: get three hist line 770)**. Using k-means clustering it was validated that the two peaks on the bimodal distributions were belonging to an active and inactive groups. On account of this, a cutoff value for each cell type was determined, which were used to binarize the AUCell scores from observed **(Figure: new\_aucell\_norm)**.

A percentage of active cells was calculated for each cell type within each TF **(Figure: aucell\_scores\_norm\_top100\_cell\_type\_binarized\_percentage)**. That being the case, the executed binarization was not conducted based on each TF as it was initially intended, but was done according to the differences within each cell type among all 100 TFs from RF. Finally, each TF was assigned to a cell type, where it had most active cells **(Sup. Table: make a table TF-Cell Type line 881)**. To conclude, 43 TFs are classified as memory, 40 as exhausted and 17 as transitional.

**Network of 100 TFs from RF: 931**

To visualize the TFs, we only used only pySCENIC interactions that appeared in 90% of runs. Further restriction of interactions to only between 100 TFs results in a network with 209 interactions and 82 TFs **(Figure: 82\_RF\_TFs)**.

Out of 82 TFs that were left from RF analysis, 38 were more active in exhausted state, 32 in memory sate and 12 in exhausted state. Out of 209 interactions 36% were confirmed by at least one of our validation sources.

Using Louvain clustering 5 TF communities were identified. Communities two, three and four have a dominant cell type presence among their TFs, community two being dominantly exhausted, meanwhile both communities three and four dominantly memory. Community one consists only of two TFs and is therefore too small to make conclusions about its belonging. Community five has a more heterogeneous composition of TFs. **LITERATURE CHECK?**

When driving comparison with Bolouri paper network, their share 9 TF them being: EOMES, EZH2, FOS, JUN, MYC, NFATC2, NFKB1, PRDM1 (BLIMP1 in Bolouri paper Network) and RUNX3. All but 2 overlapping TFs have the same cell type assigned to. The discrepancies being: EZH2 and NFATC2. **WHY IN THE DISCUSSION OR HERE?**

To reduce number of nodes and interactions to a greater extend, the network was reduced to only these 9 TFs and their direct interacting neighbors. As a consequence, this size reduction led to a network with 41 nodes and 106 interactions **(Sup Figure: 41\_paper\_overlap\_TFs)**. Two communities were identifiable after Louvain clustering, one of them being dominantly exhausted and another one dominantly memory.

**Fusion of reduced pySCENIC network and Bolouri paper network: 1213**

Prior to overlapping reduced pySCENIC network and Bolouri paper network first a nomenclature unification was not to be overseen. For example, FOS, FOSB, JUN, JUNB and JUND nodes had to be all combined in one node in order to match the JUN-FOS node in Bolouri paper network (also referred to as AP1 heterodimer). Similarly, a new node for a cooperative regulation by a protein complex NFATC1:JUN-FOS:IRF4:BATF was added, since our model does only consider TF-TF regulation and Bolouri paper network has such more complex features. Secondly, the same procedure of extracting 9 overlapping nodes and their direct interacting neighbors was be conducted.

The resulting fusion network consists of 66 nodes and 170 interactions **(Figure: 66\_bolouri\_RF\_fusion)**. 28 TFs belong to exhausted group, 27 to memory group and 2 to transitional group. Additionally, 8 cell receptors were incorporated from Bolouri paper network and the previously mentioned protein complex was manually added. Out of 170 interactions 64 came from Bolouri paper and 106 from pySCENIC network, out if which 37% are confirmed by at least one of our validation sources. The NIS score, which determines the weight of an interaction was conserved for the pySCENIC interactions and was set equal 1 for the Bolouri paper ones. Interactions between these two models do not overlap.

Louvain clustering revealed 4 TF communities. Communities one and three are major in exhausted and memory TFs respectively. Communities two and four do not show any particular dominance and are well mixed.

**Reduce the fusion network to relevant for the computation nodes**

Not all the nodes from the fusion network are important for the next step, which is determining the logical rules. Nodes that exclusively possess incoming or outgoing interactions can be discarded for the next step, since they would only change the activity of other nodes but do not have any additional information when it comes to determining logical rules to decide if another TF is on or off. A good example of such nodes can be seen in community three grouped around JUN-FOS TF **(Figure: 66\_bolouri\_RF\_fusion)**.

After this manual curation the resulting network has 38 nodes and 103 interactions **(Figure: 66\_bolouri\_RF\_fusion\_2)**. 18 exhausted TFs, 14 memory TFs, 1 transitional TF, as well as 4 cell receptors and 1 protein complex were. When it comes to interactions 34 from the paper and 107 from pySCENIC, 19% of which were confirmed. The 4 communities became self-evidently reduced; however their composition maintained the proportions corresponding to the non-curated version of the fusion network.

**Curated fusion network segmentation**

Even though, our fusion network became significantly reduced it is still too big for logical rules simulation. That being the case, specific network modules were selected to discover the rules separately, as a possible approach to overcome this complexity issue.

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

**Discussion Points**

* Cistrome peaks and extensions
* RF TFs
* Difference in overlapping nodes qualification
  + EZH2 part of a complex or subunit of PRC2?
  + NFATC2 on the borderline of activity from AUCell