**Abbreviation list:**

BN: Boolean network

CD8: cluster of differentiation 8

CTLA-4: cytotoxic T lymphocyte antigen 4

GRN: genetic regulatory network

HCV: hepatitis C virus

IS: Interaction score

MHCI: major histocompatibility complex 1

NIS: normalized interaction score

PD-1: programmed cell death protein 1

pySCENIC: python implementation of the Single-Cell Regulatory Network Inference and Clustering

RF: Random Forest

scRNA-seq single cell RNA sequencing

TCR: T-cell receptor

TF: transcription factor

**Introduction:**

**CD8+ T-cells**

Cytotoxic T-cells also known as CD8+ T-cells or T-killer cells belong to lymphocyte cell class. Originally, they develop from hematopoietic stem cells and later on relocate to thymus as T-cell progenitors, where further development stages occur, till the cells mature and differentiate cell receptors, for instance T-cell receptor (TCR) and cluster of differentiation 8 (CD8). TCR is a major player during antigen recognition. It is specific for each antigen and is located on the surface of a cytotoxic T-cells. There it interacts with major histocompatibility complex class one (MHCI) proteins, which are located on the surface of all nucleated cells. MHCI is a molecule that presents the antigen to the T-killer cells, which in its turn destroys the infected cell. This antigen presentation process is supported by CD8 glycoprotein, which acts as a coreceptor and binds to MHCI as well (Vohr, 2016).

Hence, CD8+ T-cells act antigen specifically during adaptive immune response, for example in cases of viral infection or tumors. CD8+ T-cells are characterized upon their interaction with the antigens. Naïve T-cells are the ones that have not had contact with antigen yet. Consequently, they are a precursor of effector T-cells, which are the cells that upon antigen presentation develop characteristics like cytotoxicity and ability to produce cytokines. Finally, after the antigen has been eliminated, effector T-cells differentiate into a less active, memory state. These cells can be reactivated in case there is need to eliminate the antigen to which their TCRs are specific to. Thus, a quicker immune response takes place in case of a secondary antigen exposure (Murphy, 2017).

**Chronic and acute viral infections**

Viral infections can be classified according to antigen exposure time. This exposure time has a detrimental influence on development stages of cytotoxic T-cells. A shorter exposure to an antigen, known as acute infection, CD8+ T.-cells undergo classic maturing stages from naïve to effector state and after eliminating the source of infection and target cells the few surviving effector cells transform to memory cells **(Figure a: slide 2 Vorpraktikum ppt)**. The same behavior occurs after vaccination (McLane et al., 2019).

However, during chronic infections (or cancer) the antigen exposure is more prolongated and is excessive. Epigenetic changes, such as a greater expression of inhibitory receptors like programmed cell death protein 1 (PD-1) and cytotoxic T lymphocyte antigen 4 (CTLA-4), are a result of these antigen signals. This strongly alter maturation of CD8+ T-cells **(Figure a: slide 2 Vorpraktikum ppt)**. Thereupon, T-Cells do not develop into fully functional effector or memory state. This phenomenon is known as exhaustion. This cellular state results in inability of T-Cells to fully eliminate the antigen and to differentiate into memory cells for a reoccurring infection (Kurachi, 2019).

It has been already demonstrated that the effects of the T-cell exhaustion can be reversed, therefore reviving functional CD8+ T-Cells (Barber et al., 2006). This provides grounds for investigating new therapeutic possibilities for cancer and viral infection treatment.

**Chronic hepatitis C virus infection**

Chronic hepatitis C virus (HCV) infection is a consequence of a failure in direct virus elimination and can be linked with reduced effectivity of CD8+ T-cells caused by their exhaustion (Hofmann et al., 2021). It is a liver condition caused by an RNA virus. Left unsupervised and without proper care it could lead to and chronic liver inflammation such as fibrosis, cirrhosis and hepatocellular carcinoma (Zaltron et al., 2012).

During chronic HCV infection exhausted CD8+ T-cells coexist in three states: memory-like, transient and terminally exhausted. For simplicity these states will be referred to as memory, transient and exhausted throughout this study. These states are characterized by differences in expression of genetic markers, specifically CD127, which is more expressed in memory cells. Moreover, it is demonstrated that memory cells must go through transient state to become exhausted. There are various up and down regulations, that are known to be common in exhausted cytotoxic cells. However, distinguishable signatures for each of the above-mentioned states are less clear (Hensel et al., 2021).

Direct-acting antiviral therapy treatment has been demonstrated to be highly effective against HCV infection Even though, it is possible to achieve virus elimination, an exhaustion epigenetic footprint is left (Hensel et al*.*, 2021). This highlights the need for more rigorous investigation into therapeutic approaches and a better understanding of exhaustion states. Particularly, changes in gene expression and activity, that take place during this process.

**Gene regulatory networks**

A gene regulatory network (GRN) is a collection of entities (genes, proteins or molecules) that interact with each other and describe a state and/or changes between multiple states. The network itself and the interactions between the entities are based on experimental data. Each network subunit is represented as a node that can be linked to further subunits **(Figure: example of GRN)**. A node can have multiple incoming and outcoming interactions. To determine when a node belongs to one state or another logical laws are used (Kaern et al., 2003).

The purposes of these networks can vary a lot. They could only have a descriptive character and visualize interactions within a specific cell state. Others may be used to create *in silico* knock outs or upregulations to simulation system perturbations, make predictions and guide lab experiments. Besides that, they can also provide insight into biological processes, by highlighting its architecture and dynamics, and be a method to combine empirically driven data and theory to find answers on biological questions (Karlebach and Shamir, 2008).

**Boolean Networks**

A Boolean network (BN) is a specific example of a GRN. It discards all the intermediate states by making a simplification of considering only two node sates, active (1) and inactive (0) (Ramírez and Mendoza, 2017). As the result, interactions, that determine the state of each node, are also simplified, to strictly being activating or inhibitory. The binarization of nodes scores is usually performed by setting a threshold. Expression or activity values (depending on the network) above the threshold are set to 1, while the ones below to 0 (Ristevski, 2015).

Models created by BNs are relatively simple and reduce the network dynamics with the state binarization. Therefore, they are less sensitive to slight gene activity or expression changes. They are also more suited for relatively small networks, as the number of states for such network grows exponentially: 2n (n being the number of nodes in the network) (Ristevski, 2015).

**Predicting TF activity**

Python implementation of the Single-Cell Regulatory Network Inference and Clustering (*pySCENIC*) was implemented to obtain targets of the transcription factors (TFs), that are present in CD8+ T-cells. *pySCENIC* is a tool that works with single cell RNA sequencing (scRNA-seq) data and represents a combination of 3 other tools, that analyze the input in a stepwise manner (Aibar et al., 2017).

In the first step,*GENIE3* or *GRNBoost*, is used to determine co-expression between genes from scRNA-seq data, creating a regulon of targets for each TF. Each identified interaction is given an importance score (Van de Sande et al., 2020). However, some of this co-expression may be result of not *cis*-regulatory elements, that directly interact with neighboring genes, but of *trans*-regulatory elements, which act on more distant genes. For that reason, *RcisTarget* is implemented to reduce regulons and keep only *cis*-interactions (direct targets). This is achieved by conducting enrichment analysis of TF binding sequence and the target gene location. Finally, activity of each TF is quantified in each cell by *AUCell* (Aibar *et al.*, 2017).

One *pySCENIC* run consists of these 3 steps, For the purpose of this project *pySCENIC* was iterated 50 times to obtain more robust regulons and interactions.

**Objective and premise of the thesis**

The focus of this project lies in investigating chronic viral infection caused by HCV and its effects on the CD8+ T-cell activity. The objective is to create a GRN to describe CD8+ T-cell exhaustion, considering HCV effects on gene expression. This GRN should describe interactions among TFs and bring forward a coherent outline of processes on the intracellular level, that take place during the exhaustion. The ultimate goal is to further comprehend and represent the genetic differences between the three cell states (memory, transient and exhausted).

Two other studies are used as a reference in this project. Bolouri et al paper has already created a network to describe CD8+ T-cell exhaustion based on the avaliable literature sources and information from other models **(Figure: Bolouri Network)**. The Hérault et al paper has computed a BN to describe early hematopoiesis aging, baed on scRNA-seq data and *pySCENIC* implementation (Hérault et al., 2022).

This strudy aims to use computationally obtained regulons to create a GRN and curate this network with data from Bolouri et al paper, thus using it as literature source for higher integrity. The workflow and idea behind this project are analogous to the Hérault et al paper. Nonetheless, they are applied on a different problem with its unique hurdles.

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

To run *pySCENIC* an input table with gene expression in each cell is required. From the raw scRNA-seq data only expression information in cells that had a cell type assigned was kept. There are three different cell types: memory, transient 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*** **clean up: 173**

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.

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, generated during the first step of *pySCENIC* run, 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 transient.

**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 transient 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 transient 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
* Segmentate Bonesis network and explain the segments

# Works Cited

Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. Nature Methods *14*, 1083-1086. 10.1038/nmeth.4463.

Barber, D.L., Wherry, E.J., Masopust, D., Zhu, B., Allison, J.P., Sharpe, A.H., Freeman, G.J., and Ahmed, R. (2006). Restoring function in exhausted CD8 T cells during chronic viral infection. Nature *439*, 682-687. 10.1038/nature04444.

Hensel, N., Gu, Z., Sagar, Wieland, D., Jechow, K., Kemming, J., Llewellyn-Lacey, S., Gostick, E., Sogukpinar, O., Emmerich, F., et al. (2021). Memory-like HCV-specific CD8+ T cells retain a molecular scar after cure of chronic HCV infection. Nature Immunology *22*, 229-239. 10.1038/s41590-020-00817-w.

Hérault, L., Poplineau, M., Duprez, E., and Remy, É. (2022). A novel Boolean network inference strategy to model early hematopoiesis aging. bioRxiv. 10.1101/2022.02.08.479548.

Hofmann, M., Tauber, C., Hensel, N., and Thimme, R. (2021). CD8(+) T Cell Responses during HCV Infection and HCC. J Clin Med *10*. 10.3390/jcm10050991.

Kaern, M., Blake, W.J., and Collins, J.J. (2003). The engineering of gene regulatory networks. Annu Rev Biomed Eng *5*, 179-206. 10.1146/annurev.bioeng.5.040202.121553.

Karlebach, G., and Shamir, R. (2008). Modelling and analysis of gene regulatory networks. Nature Reviews Molecular Cell Biology *9*, 770-780. 10.1038/nrm2503.

Kurachi, M. (2019). CD8+ T cell exhaustion. Seminars in Immunopathology *41*, 327-337. 10.1007/s00281-019-00744-5.

McLane, L.M., Abdel-Hakeem, M.S., and Wherry, E.J. (2019). CD8 T Cell Exhaustion During Chronic Viral Infection and Cancer. Annual Review of Immunology *37*, 457-495. 10.1146/annurev-immunol-041015-055318.

Murphy, K.W.C.J.C. (2017). Janeway's immunobiology.

Ramírez, C., and Mendoza, L. (2017). Phenotypic stability and plasticity in GMP-derived cells as determined by their underlying regulatory network. Bioinformatics *34*, 1174-1182. 10.1093/bioinformatics/btx736.

Ristevski, B. (2015). Chapter Three - Overview of Computational Approaches for Inference of MicroRNA-Mediated and Gene Regulatory Networks. In Advances in Computers, A.M. Memon, ed. (Elsevier), pp. 111-145.

Van de Sande, B., Flerin, C., Davie, K., De Waegeneer, M., Hulselmans, G., Aibar, S., Seurinck, R., Saelens, W., Cannoodt, R., Rouchon, Q., et al. (2020). A scalable SCENIC workflow for single-cell gene regulatory network analysis. Nature Protocols *15*, 2247-2276. 10.1038/s41596-020-0336-2.

Vohr, H.-W. (2016). Encyclopedia of immunotoxicology.

Zaltron, S., Spinetti, A., Biasi, L., Baiguera, C., and Castelli, F. (2012). Chronic HCV infection: epidemiological and clinical relevance. BMC Infect Dis *12 Suppl 2*, S2. 10.1186/1471-2334-12-s2-s2.