**Acknowledgements**

To begin with I would like to express my deepest gratitude to PD Dr. Carl Herrmann for giving me an opportunity and his trust to work on this project. His lectures in Data Analysis and Bioinformatics have inspired me to do my Bachelor’s thesis in this filed. I’m extremely grateful to Dr. Carlos Ramírez, who has been my contact person for all possible issues. Discussions with him have always been amusing and provided new insights and ideas how to approach the project. Special thanks to the creators of *Bonesis*, S. Chevalier and L. Paulevé for instantly answering and solving my problem with the tool. Additionally, I would love to say thank you to the entire HDSU team, who have also assisted me with my struggles and provided an amazing working atmosphere. Lastly, I want to thank my parents and close friends, who have morally supported me during this time.

**Abbreviation list:**

AUC: area under the curve

BN: Boolean network

CD8: cluster of differentiation 8

CTLA-4: cytotoxic T lymphocyte antigen 4

DAA: direct-acting antiviral therapy

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 : 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 (DAA) 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 calculating the area under the curve (AUC) using *AUCell* (Aibar *et al.*, 2017).

**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.

**Materials**

**Data Description**

The entire data used during this project is provided by University Hospital Freiburg, Faculty of Medicine. **ANY GROUP IN PARTICULAR I SHOULD REFER TO?**

The raw scRNA-seq data of the CD8+ T-cells contains count values of 43 779 genes among 1 258 cells. ATAC-seq analysis results of 4 676 genes were provided. Additionally, cell state annotations to 784 were contributed.

**Documentation**

The complete code for this project can be accessed on <https://github.com/IlyaSchneider10/GRN_Tcell_Exh>. A more detailed file description is provided there. The input data, figures and intermediate results are accessible on the *curry0* cluster of BioQuant Heidelberg (curry0.bioquant.uni-heidelberg.de) under */media/ag-cherrmann/ischneider/GRNTcellExh*.

**Software: make a table with links?**

*R* v4.0.2

*Python* v 3.10.4

*pySCENIC* v0.10.0 (Van de Sande et al., 2020)

*Bonesis* v0.0a0 (<https://github.com/bioasp/bonesis>)

Cytoscape v3.9.1

GenomicFeatures, R package v1.42.3

GenomicRanges, R package v1.42.0

TxDb.Hsapiens.UCSC.hg38.knownGene, R package v 3.10.0

Homo.sapiens , R package v1.3.1

plyranges, R package v1.10.0

AUCell, R package v1.12.0

GSEABase, R package v1.52.1

GEOquery, R package v1.52.0

openxlsx, R package v4.2.5

stringr, R package v1.4.0

gplots, R package v3.1.3

igraph, R package v1.3.2

RColorBrewer, R package v1.1.3

ComplexHeatmap, R package v2.6.2

randomForest, R package v4.6.14

mousetrap, R package v3.2.1

LaplacesDemon, R package v16.1.6

devtools, R package v2.3.2

dorothea, R package v1.2.2

Seurat, R package v4.1.1

ggpubr, R package v0.4.0

ggplot2, R package v3.3.6

**Methods**

**Data Preprocessing for *pySCENIC* pipeline:**

From the raw scRNA-seq data only expression information in cells that had one of the three cell states (memory, transient, exhausted) assigned was kept. Following that, 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\*number of cells**

Some differences in nomenclature between our scRNA-seq data and Boulori paper network were identified and replaced. Out of 57 nodes, 46 were already added during the gene filtering step. To maximize the amount of Bolouri et al genes in the *pySCENIC* input 7 genes that did not meet filtering criteria were added manually afterwards. As a result, a table containing expression data of 10 245 genes among 784 cells was used as *pySCENIC* input.

**Transcription Factor Activity Prediction:**

One *pySCENIC* run consists of these 3 steps, described in detail in the introduction section. For the purpose of this project *pySCENIC* was iterated 50 times to obtain more robust regulons and interactions. All the TFs that were given AUC score by AUCell were kept as reference TFs. Therefore, 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 from the TF reference list.

To reduce the scope of interactions to be analyzed even further, only interactions that appeared in 40 or more *pySCENIC* runs (80% of the runs) were kept. 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 by *GRNBoost*, 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. It is a mean to weight the interactions and to identify communities inside of a network.

**Verification of the interactions: 373**

To validate *pySCENIC* interaction databases *Cistrome* and *DoRothEA* were chosen. Additionally, ATAC-seq data from patients who had hepatitis C viral infection and were treated with DAA was used as another source of confirmation.

*Cistrome* is based around ChIP-seq experimental data. Only bed files with TFs that were classified as “Blood” in the tissue type category were selected for further analysis. For each TF peaks were arranged according to their score and 6000 peaks with the highest values were kept. Each of the peaks was expanded 1 kb in both directions in order to overlap it with human genome. The genes that were identified in this TF extended window were considered its targets. *TxDb.Hsapiens.UCSC.hg38.knownGene* was used as reference for transcripts. To conduct the overlap between extended peaks and the transcripts *intersect()* from *GenomicRanges* was implemented. Finally, to get genes for the overlapped transcripts, package *Homo.sapiens* was utilized as reference genome and *find\_overlaps()* from *plyranges* was applied to extract gene names.

*DoRothEA* database is supported on transcriptional data and has a confidence score for every interaction from E to A, E being the lowest score and A the highest. Out of all overlapping interactions only the ones that had A confidence were kept.

ATAC-seq data originates from the same cells as the scRNA-seq data and thereby is of high relevance. Open chromatin regions that were identified were mapped with TF and compared with *pySCENIC* interactions. The overlap was marked as confirmed.

**Network modeling**

TFs that were left after 80% frequency cut were further reduced using Random Forest (RF) approach. The idea was to use AUC scores from TFs as explanatory variables to explain the belonging of the cells to one or another cell type (response variable). TFs were ranked according to their importance scores and top 100 were kept for the network modulation: *randomForest( as.factor(cell\_type) ~ ., data = random\_forest\_input, importance = TRUE)*

Interactions between these 100 TFs were selected from the ones retained after 80% filtering and further reduced to only interactions that appeared in at least 90% of *pySCENIC* runs.

**Logical Rules Formulation**

To study network dynamics and logical rules behind it we implemented *Bonesis* (<https://github.com/bioasp/bonesis>). *Bonesis* was used as the tool to establish rules under which nodes interact with each other and switch between cellular states. The main premise was that cells must start in the memory state and go through transient configuration to finally end exhausted. Exhausted state was additionally defined as a fixed point, meaning that once the network reached this state it gets trapped in it and cannot switch to other states. Consequently, the attractor states describe T-cell exhaustion. Besides definition of state transitions, this tool also requires characteristic information about each of them. After normalizing and scaling our raw scRNA-seq data using *Seurat* package, gene expression data was used to create cellular sate profiles

**Results**

***pySCENIC*** **Output Analysis**

In total, *pySCENIC* identified 353 different TFs among 50 runs, with 6 254 unique interactions. After the reduction to the most robust interactions, present in 80% of the runs, 641 interactions among 251 TFs were left.

29% (73 TFs) of these TFs were found in the Cistrome database subset were selected. In total 20% of interactions that made it through the frequency filter were confirmed. Dorothea contributed to confirming 2% of most frequent interactions. It is mportant to note, that none of these interactions were identified via other two validation methods that were used. ATAC-Seq data verified 15% of interactions, that passed the frequency filter.

To, conclude, 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. Full lists of targets (not only restricting TF-TF interactions) were used as input regulons for each TF. 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 did 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 states within a TF could be observed **(Sup Figure: new\_aucell\_norm)**.

After reducing absolute amount of TFs to 100 using RF 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. That being the case, a cutoff value for each cell type was determined, which were used to binarize the AUCell scores from **(Figure: new\_aucell\_norm)**.

A percentage of active cells was calculated within each cell state for every TF **(Figure: aucell\_scores\_norm\_top100\_cell\_type\_binarized\_percentage)**. 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:**

To visualize the TFs, interactions were reduced to the once that appeared in 90% of *pySCENIC* 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 and NIS scores 5 TF communities were identified. Communities 2, 3 and 4 have a dominant cell type presence among their TFs, community 2 being dominantly exhausted, meanwhile both communities 3 and 4 dominantly memory. Community 1 consists only of two TFs and is therefore too small to make conclusions about its belonging. Community 5 has a more heterogeneous composition of TFs. **LITERATURE CHECK?**

When driving comparison with Bolouri paper network, *pySCENIC* network shares 9 TFs with it, 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, in *pySCENIC* network are assigned to exhausted cell state. **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:**

Prior to overlapping reduced *pySCENIC* network and Bolouri paper network a nomenclature unification was not to be overseen. For example, *FOS*, *FOSB*, *JUN*, *JUNB* and *JUND* nodes had to be all combined into one node in order to match the *AP1* heterodimer node in Bolouri paper network (in the network appears as *JUN-FOS*). 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 takes into more complex features, like this cooperative action of these four genes. Furthermore, the same procedure of extracting 9 overlapping nodes and their direct interacting neighbors was 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 of 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 to 1 for the Bolouri paper ones. Interactions between these two models do not overlap.

Louvain clustering revealed 4 TF communities. Communities 1 and 3 are dominantly exhausted and memory-like respectively. Communities 1 and 4 do not show any homogeneity.

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

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

After this manual curation the resulting network has 38 nodes and 103 interactions, 18 exhausted TFs, 14 memory TFs, 1 transient TF, as well as 4 cell receptors and 1 protein complex **(Figure: 66\_bolouri\_RF\_fusion\_2)**. 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.

**Further network reduction**

Even though, the fusion network became significantly reduced it is still too complex to define logical rules. That being the case, we adjusted current network by drastically simplifying it. First, exclusively interactions from Bolouri et al paper were kept. Second, as in previous network simplification, only nodes that are relevant for logical rules were kept. The resulting network contains 13 nodes with 23 interactions among them **(Figure: 13\_Bonesis\_toy)**.

**Establishing logical rules**

The size of this network was suitable to begin analyzing it with *Bonesis*. A dotplot of gene expression from the *pySCENIC* input data was computed to create cellular sate profiles **(Figure: Dot\_plot)**. Based on this plot and the TF-activity heatmap **Figure: aucell\_scores\_norm\_top100\_cell\_type\_binarized\_percentage)** cellular configurations for each state were created **(Table: bonesis 2.1.1)**. As the result, further nodes have adopted a different state, when compared to Bolouri et al. paper. *NFATC1* describe memory state, meanwhile *FOXO1* switched to exhausted cell type. Therefore, the colors of the nodes had to be adjusted **(Figure: 13\_Bonesis\_toy\_adjusted)**.

Ultimately, after conduction simulation for logical rules and reducing solutions to the ones that can reach exhaustion from the memory state 355 806 networks that fulfill such constrains were obtained. The following table represents a summary of unique rules, that describe activity of each node **(Table: proj 2.1.1)**. Nodes for *TCRs* (*CD3, TCR, CD8, CD28*), *CTLA-4* and *PD-1* do not have any logical solutions and are shown to be constantly turned on during the exhaustion, suggesting that these nodes do not have enough inputs to describe changes in their activity. Consequently, this model in total appears to be too simple to describe nodes’ activity, so further modifications are required.

**Splitting the model into two regions**

As the scalability of *Bonesis* is limited in number of nodes and interactions it takes in. The top half, consisting of 6 nodes (*NFATC1*, *CTLA-4*, *PD-1*, *NFATC1:JUN-FOS:IRF4:BATF* and *EZH2*) and the bottom half with 7 nodes (*TCF-1*, *BCL6*, *BACH2*, *PRDM1*, *ID3*, *IL2-R* and *FOXO1*) **(Figure: make one)**. The split is justified by the architecture of the network since the top half does not receive any input signals from the bottom part, therefore being the driving portion of the network. **BOLOURI FOR FURTHER REASONS**.

The idea behind this approach is to look for informative input nodes and interactions to explain the activity switches of the three nodes from the top half (TCRs, CTLA-4 and PD-1), without drastically increasing running time and amount of solutions provided by *Bonesis*.

**Additional nodes from curated fusion network.**

Reduced fusion network **(Figure: 66\_bolouri\_RF\_fusion)** was scanned for the interactions, that would have one of three problematic nodes as a target. The goal was to add minimum number of TFs, while also try and include some interactions from *pySCENIC*.

As an additional input to the *CD3, TCR, CD8, CD28* node *BTLA* was introduced. In order to connect *BTLA* with the rest of the network *NFKB1* and *JUN-FOS* were incorporated. Interaction between *JUN-FOS* and *NFKB1*, as well as between *EZH2* and *JUN-FOS* originate from *pySCENIC* and have no description of the interaction being activating or inhibiting. To determine that, the activities of these TFs were compared in each cell state using the heatmap **(Figure: aucell\_scores\_norm\_top100\_cell\_type\_binarized\_percentage)**. Interaction from *JUN-FOS* was defined as activating, meantime from *EZH2* to *JUN-FOS* inhibiting.

*NFATC2* was selected to address *PD-1* and *CTLA-4*. As in the previous case, by incorporating only this node it would have no input interactions and subsequently it is not possible to switch it on. *STAT1* was picked as it targets *NFATC2* and it is targeted by *EZH2*. The sign of the interaction between these nodes was determined as formerly with interactions originating from *pySCENIC*, in this case it being activated. By adding *NFATC2* to this network an activating interaction to *NFATC1*.

The resulting network consists of 11 nodes and 17 interactions **(Figure: bonesis\_final\_3)**. It was tested to determine the logical rules and verify that the issue with the nodes having one fixed state solution was eliminated. The cellular profiles on the **(Figure: bonesis\_final\_3)** were determined as earlier using a dot plot **(Figure:dot plot 2)**. 25 792 networks that could describe transition from memory to exhausted state were identified.

The bottom part of the split network was brought back and incorporated with the expanded top part **(Sup Figure: make a full network in bonesis)**. However, after performing Bonesis simulation the increase in number of solutions was so drastic, that it was not possible to exactly determine it. The inclusion of only 2 nodes of the 7 bottom nodes, *BCL6* and *TCF-1*, mounted the total amount of solutions networks to 1 049 908. This output was considered too big and not analyzable. For that reason, it was decided to continue the project with the network from **(Figure:dot plot 2)**, since its output is more manageable and it fulfills the criteria of having at least one conditional solution for each node, which was not the fact with the previously split network from **(Figure: 13\_Bonesis\_toy)**.

**Analyze network’s solutions**

**(Table: final\_3)** represents exhibits unique rules for each node and their share within every node. Even though, all nodes posses a conditional solution, all of them also have a fixed value (0, 1 or both) as well.

For *TCRs*, *CTLA-4*, *NFATC1:JUN:FOS:IRF4:BATF*, *NFATC2*, *PD-1* and *STAT1* fixed value/values represented 50% or less of the total node solutions. Thereupon, the solution networks containing these values in these nodes were removed, leaving a total of 1 381 networks.

The solution percentages after this curation are represented in the **(Table: final\_3\_curated\_1)**. The same procedure was done for *BTLA*, *EZH2*, *JUN-FOS* and *NFKB1*, since they now also qualify according to the former criteria. 64 networks were left. The final set of rules for each node are depicted in **(Table: final\_3\_curated\_2)**.

For these 64 solutions the attractor states were extracted from the *Bonesis* output **(Table: final\_3\_curated\_2\_attractors)**.

# 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.