In silico identification of Yap/Taz bound transcription factors in the lung epithelium of Idiopathic Pulmonary Fibrosis

TRABAJO FIN DE GRADO

Laura Martínez Rábade

Grado en Biomedicina

Facultad de Ciencias Experimentales

UNIVERSIDAD FRANCISCO DE VITORIA

WAGNER LAB-LUNG BIOENGINEERING AND REGENERATION

Supervisión: Beatriz González Gálvez, Hani N. Alsafadi

Junio 2021

ACKNOWLEDGEMENTS

First, I would like to express my special thanks of gratitude to Lund University and the Lung Bioengineering and Regeneration (LBR) Lab for giving me the opportunity to do research here in Lund and hosting me so kindly, but especially to Hani and Darcy for teaching me during these months and for supporting and helping me with my project.

I want also to thank my family and friends for always being supportive along the way and making me feel like home even thousand miles away.

Finally, I would like to thank my tutor Beatriz, not only for supervising this project but for always being that kind to me and being willing to help me with anything.

TABLE OF CONTENTS

1. ABSTRACT

Idiopathic pulmonary fibrosis (IPF) is the most common type of the idiopathic interstitial pneumonias. It is characterized by a progressive respiratory failure due to accumulation of extracellular matrix (ECM) in distal lung tissue and is proposed to be initiated due to recurrent injury to the epithelium and subsequent secretion of ECM components by fibroblasts. IPF is an incurable disease, and the current developed medications can only slow disease progression. Several studies have previously shown that in IPF, the Hippo signalling pathway and its main effectors YAP/TAZ, are dysregulated. In the absence of inhibitory phosphorylation by the Hippo kinases, YAP/TAZ translocate to the nucleus, where they interact with other transcriptional (co)factors to regulate target genes. This controls a wide range of homeostatic biological processes that are active especially during development such as cell differentiation, proliferation, migration, and organ size. It is known that the activity of YAP/TAZ is increased in IPF, but the specific transcription factors (TFs) binding partners are not known yet in this context.

Using Cleavage Under Targets and Release Using Nuclease (CUT&RUN), transcriptional motifs of genes targeted by transcriptional complexes containing Yap/Taz were identified in proximal and distal lung epithelial cells. The aim of this thesis is to develop, test and implement computational pipelines to identify putative YAP/TAZ co-transcriptional factor binding partners using state of the art bioinformatics tools. The developed workflow consists of motif analysis and enrichment analysis leading to a selection of TFs as potential candidates. Subsequently, validation against publicly available single-cell dataset was performed to confirm their role in a cell type specific context through the generation of TF specific signature scores.

After optimizing our analysis pipeline, we performed it on a dataset generated from CUT&RUN performed on proximal epithelial cells (i.e. isolated from murine trachea) and we identified 28 potential TFs with suitable characteristics for future experimental validation in the laboratory. Interestingly, projecting signature scores on a single cell dataset revealed that some TF signature is present only in proximal cells such as FEV, while others are present across the whole epithelial cellular landscape despite that the processed data was obtained from proximal cells. Since YAP/TAZ are involved in several pro-regenerative and homeostatic processes as well as implicated in disease pathogenesis, the identification of pathologic TFs interactions causing the disease could potentially help for defining therapeutical targets while maintaining the useful YAP/TAZ functions of repair-promoting interactions.

Keywords: Idiopathic pulmonary fibrosis, bioinformatics, transcription factors, Hippo pathway, single-cell RNA sequencing.

2. INTRODUCTION

2.1 Idiopathic pulmonary fibrosis (IPF)

Idiopathic pulmonary fibrosis (IPF) is a chronic lung disease of unknown etiology consisting in a decline in lung function, progressive respiratory failure due to accumulation of extracellular matrix in the distal lung, and high mortality (1). It is the most common of the idiopathic interstitial pneumonias, with an average life expectancy between 2 and 3 years from the time of diagnosis, yet some patients live longer. Although several individual clinical variables are correlated with survival, there is no effective method to combine these predictors, so IPF has a poor prognosis (2).

Furthermore, despite the fact that there are some recently developed medications with the ability to slow its progression (3), nowadays no effective treatments are available for this incurable disease.

2.2 The lung epithelium and IPF

The lung is composed of the proximal and distal regions, containing epithelial cells specialized for different functions (**Figure 1**). From the trachea, the conducting airways branch into bronchi and bronchioles, opening into air sacs in the most distal part, known as alveoli (4).

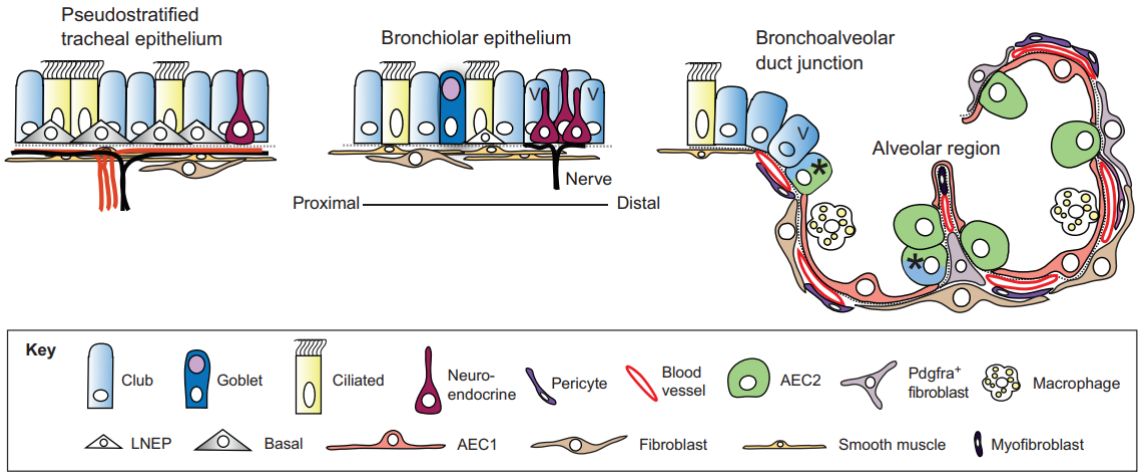

Figure 1. Epithelial cell types of the normal mouse lung. The two major epithelial cell types in the alveolar region are type II (AEC2) and type I (AEC1) alveolar epithelial cells. Source (4).

This alveolar region is where IPF is originated, playing a relevant role in this disease as recent evidence shows (3,5). There are two major epithelial cell types in the alveolus, type 1 alveolar epithelial cells (AEC1s or AT1 cells), involved in the process of gas exchange, and type 2 cells (AEC2s or AT2 cells), which maintain the homeostasis by secreting pulmonary surfactant and are also known as the progenitor cell of the alveolar epithelium (6). Moreover, there are different kinds of mesenchymal cells that reside in close proximity to the alveolar epithelial cells to help in the repair process and provide structural support.

Several epithelial progenitor cell types have been identified with the capacity for both self-renewal and replacement of a variety of specialized lung epithelial cells. In large airways and submucosal glands, basal cells are able to self-renew and differentiate into ciliated and secretory cells. In smaller airways, club cells are able to self-renew but differentiate only into club cells or ciliated cells. Also in the smaller airways, neuroendocrine cells can self-renew and give rise to club cells and ciliated cells after injury. The most limited lung epithelial progenitor cells are AT2 cells (7). They occasionally serve as alveolar stem cells with the potential to self-renew and differentiate into AT1 cells in damaged alveoli. This role of AT2 cells and basal cells as progenitors cells is critical for the regeneration of the respiratory epithelium after acute and chronic injury (8).

In IPF, fibrotic lesions and honeycomb structures replace alveolar structures of the lung after chronic injury and thus these processes fail to repair the distal respiratory epithelium so that it can still participate in gas exchange (1). There is a loss of AT1 and AT2 cells, and in addition there is also a presence of atypical epithelial cells expressing differentiated cell markers specific of proximal airways and submucosal glands in the normal lung, including basal, goblet and ciliated cell markers (1,8).

The abnormal alveolar epithelial cell homeostasis in the pathogenesis of IPF may be explained by genetic mutations affecting AT2 cells functions or survival of injured cell types which subsequently develop a pro-fibrotic phenotype. However, the contributions and responses of individual cell types to the pathogenesis of IPF and the mechanisms for the failure of alveolar epithelial regeneration are not clear yet.

2.3 The Hippo signaling pathway

The Hippo pathway is a conserved signalling pathway with many different roles in organ development, epithelial homeostasis, tissue regeneration, wound healing or immune modulation (9). The main drivers of this pathway are Yes-associated protein 1 (YAP) and transcriptional coactivator with PDZ-binding motif (TAZ), two downstream effectors that have been linked recently to the pathophysiology of fibrosis (10).

The Hippo pathway consists in a cascade of kinase complexes serving as inhibitors (**Figure 2**). When the Hippo kinase complex is active, YAP and TAZ factors are phosphorylated and subsequently sequestered in the cytoplasm or marked for degradation. This cytoplasmic localization promotes the inhibition of YAP/TAZ transcriptional activity (10).

Otherwise, in the absence of inhibitory phosphorylation by the Hippo kinases, YAP/TAZ translocate to the nucleus, where they do not bind directly to DNA but interact with other transcriptional factors to regulate target genes. The best-characterized TFs regulated by YAP/TAZ are TEAD transcription factors, composed of four members in mammals (TEAD1–TEAD4) (9).

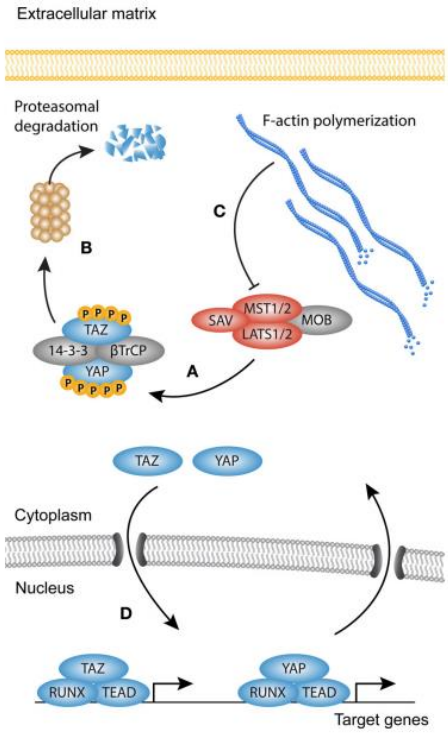

Figure 2. YAP and TAZ activation and inhibition signaling. Hypo-phosphorylated YAP and TAZ accumulate in the nucleus, where they can bind to various transcription factors, including the TEAD family, to direct gene expression changes that control a range of biological events. Source (10).

YAP and TAZ have similar mechanism of activation, but they can bind different TFs or similar TFs, as seen with the association with TEAD transcription factors. This suggests that their functions only partially overlap and share redundancy in some biological contexts.

Nuclear YAP/TAZ direct gene expression that controls a wide range of homeostatic biological processes such as cell differentiation, proliferation, cell fate decisions and organ size. They are also known for their role as mechanotransducers and mechanoeffectors. These mechanical properties of YAP and TAZ were recently translated to myofibroblast activation and the induction of fibrosis. In IPF, protein levels of both YAP and TAZ are elevated and display a predominantly nuclear localization, suggesting an increased transcriptional activity (10).

The pro-fibrotic activity of YAP/TAZ in the lung have been mainly associated with (myo)fibroblasts. However, recent studies and evidence from the Wagner group suggest an important role for YAP/TAZ in epithelial cells in the modulation of fibrotic changes. YAP and TAZ protein expression were found to be elevated in both proximal and distal lung epithelial cells (unpublished data). However, the exact role of YAP and TAZ in these regions remain unclear. Additionally, the exact transcriptional complexes that contain YAP and TAZ remain unknown in lung epithelial cells in the context of fibrosis.

Thus, it is important to continue investigating potential YAP/TAZ binding partners and identify possible specific changes between normal and diseased lung epithelium. Since YAP and TAZ interactions are context dependant, it makes it difficult to broadly target YAP/TAZ as a therapy to treat IPF as they may also be involved in pro-regenerative processes which are necessary for restoration or regeneration of normal lung tissue. Identification of exact TF interactions with YAP/TAZ, will allow for targeting the pro-fibrotic and disease-causing interactions while allowing pro-regenerative combinations to function.

To address this, a chromatin immunoprecipitation-based study targeting the motif sequences of all YAP/TAZ bound transcriptional complexes was performed at the host lab using cleavage under targets and release using nuclease (CUT&RUN). This method works by targeting the TF complexes using magnetic bead conjugated antibodies specific for YAP/TAZ and the cleavage of DNA around the TF complexes using a bacterial nuclease. The resulting cleaved complex/DNA is then immunoprecipitated and the DNA is released from this complex to be sequenced (11). This resulted in a list of motifs that are targeted by YAP/TAZ. However, a bioinformatics pipeline needs to be established to identify the bound TFs.

2.4 Objectives

The aim of this thesis is to implement the computational pipelines to identify YAP/TAZ co-transcriptional factor binding partners in the lung epithelium of IPF using state of the art bioinformatics tools. To achieve this, a workflow was developed to first select a number of TFs as potential candidates using predictive research tools and then validate them to confirm their role in a cell type specific context, making these final TFs suitable for future experimental validation in the laboratory.

3. MATERIALS AND METHODS

3.1 Identification of TFs

3.1.1 Data source

Transcription factors bind to specific DNA sequences called transcription factor binding sites (TFBSs) within promoter and enhancer regions of genomic DNA to activate or repress gene expression. These interactions can be determined experimentally with different techniques.

CUT&RUN is an epigenomic profiling strategy in which micrococcal nucleases control antibody-targeted cleavage and this releases specific protein–DNA complexes into the supernatant for paired-end DNA sequencing.

Chromatin Immunoprecipitation sequencing (ChIP-seq) has been the predominant method of mapping protein–DNA interactions for several years. However, the CUT&RUN method has advantages since it achieves better results in resolution and sensitivity, providing high-quality data even when low cell numbers are available (starting with only 1,000 cells for a transcription factor). The procedure is simpler and shorter than ChIP-seq and does not require chromatin fragmentation or solubilization (11).

Initially for this project, transcriptional motifs of genes targeted by transcriptional complexes containing Yap/Taz were identified in proximal (tracheal) and distal (alveolar) epithelial cells using CUT&RUN. Our bioinformatic analysis started with 3 datasets containing annotated gene sets for YAP, TAZ and both YAP/TAZ, obtained from the CUT&RUN experiments with tracheal epithelial cells.

3.1.2 Motif analysis and enrichment analysis

The large number of TFs that plays a role in regulation of gene expression can be a challenge when it comes to determining which of these are likely to be controlling a set of genes in specific signaling pathways. However, nowadays this can be assisted by computational prediction, utilising experimentally verified binding site motifs.

CiiiDER is a recently developed bioinformatic tool for transcription factor binding analysis. It is a downloadable program, independent of computer operating system, which can be run on a local computer and to have saved projects that can easily be revisited. This makes it very useful for a wide audience since no coding and programming skills are needed (12).

CiiiDER permits to develop two types of analysis, the **scan analysis** to identify potential TFBSs in regulatory regions, and the **enrichment analysis** to identify the over- or under-represented TFBSs compared to a user-specified background list. The first step is to paste the annotated gene list from CUT&RUN experiments in the text box and then press the button “Run Scan” to begin the site prediction for the chosen gene list (**Figure 3**). The “Run Enrichment” button also can be used to queue an enrichment analysis immediately after the scan analysis.

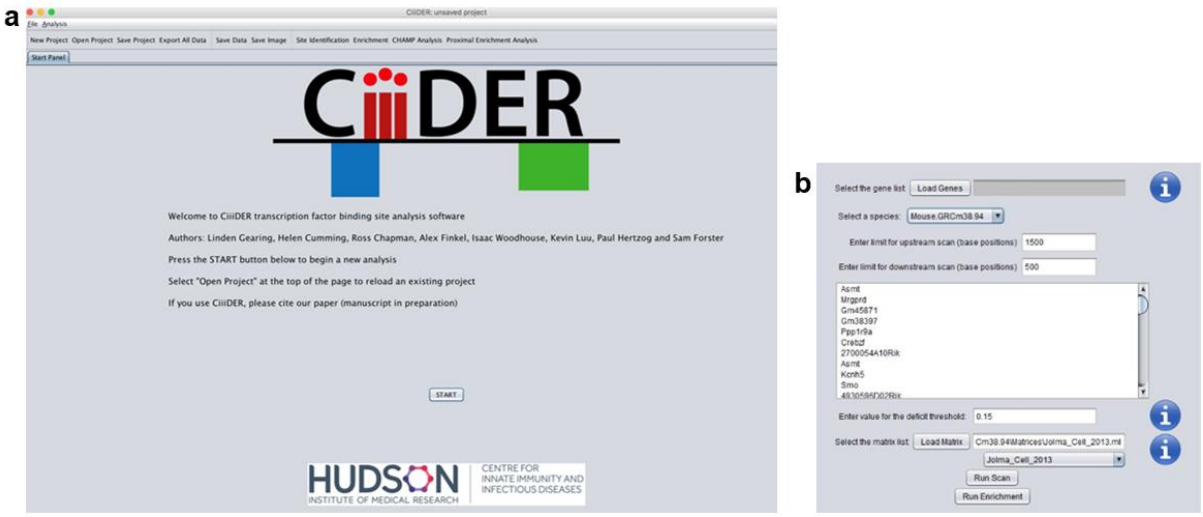

Figure 3. CiiiDER interface. (**a**) The start panel. (**b**) Scan load box.

Once the scan analysis is completed, the promoter panel with all the predicted TFBSs is displayed. On the right side of the panel there are different parameters that can be changed to explore and compare data (**Figure 4**).

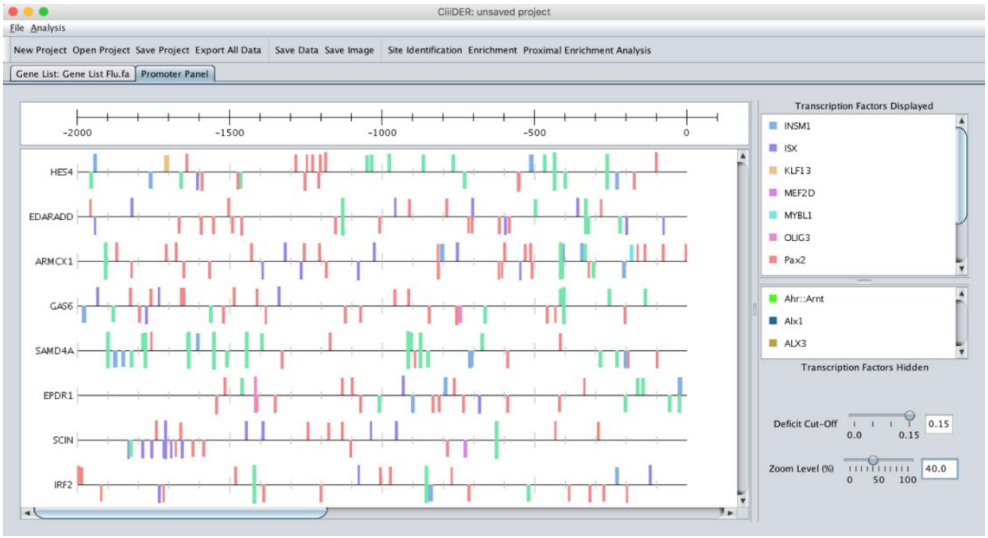

Figure 4. Interactive promoter panel interface. It shows all the potential TFBSs.

To begin the enrichment analysis, a background gene list must be provided. The selection of this gene set is important as it can alter the results that CiiiDER provides. For our analysis, this list contains a large number of highly conserved genes expressed in lung epithelial cells and not effected by a fibrotic injury.

The results of the enrichment analysis can be visualised on an interactive plot with the over- or under-represented transcription factors (**Figure 5**) and additional information about the enrichment panel can be saved in CSV format.

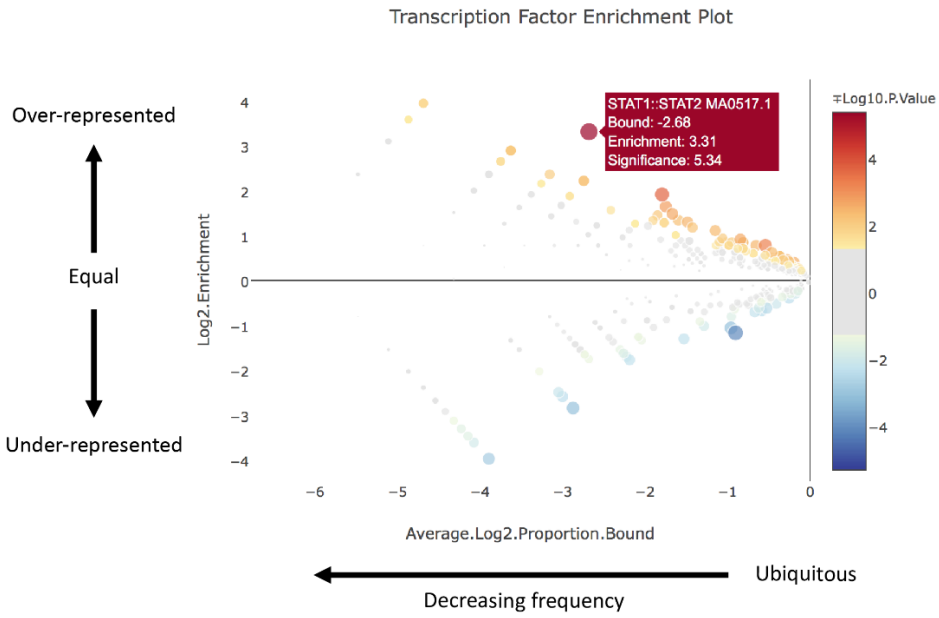

Figure 5. Interactive enrichment scatter plot. It shows over- or under-representation of each TF in the Y axis (Enrichment) and the proportion of genes that contain binding sites in the X axis (Average proportion bound). Size and colours of TFs represent their P-values (Significance Score). Source (13).

Significantly over- and under-represented TFs are calculated by CiiiDER using statistical methods to compare the numbers of sequences with predicted TFBSs to the number of those without. The TFs with over-represented TFBSs in a set of co-expressed genes are more likely to be involved in regulating the expression of these genes.

3.1.3 TFs selection

The previous analysis concludes with the identification of enriched TFs in co-expressed gene sets for YAP, TAZ and both YAP/TAZ. To be able to do the subsequent validation, a small group with the most remarkable TFs must be selected and this choice is made based on the

Significance Scores. A Significance Score cut off was selected based on the distribution of the data: a histogram representing these Significance Scores was generated in R Studio using the code below:

```
library(ggplot2)
yap_taz_data <- read.csv("TC_YAPTAZ_enrichment_data_MostSigDeficit.csv")
ggplot(data = yap_taz_data, aes(Significance.Score)) +
  geom_histogram(binwidth = 3, color = "Black", fill = "Light Blue") +
  theme_classic() +
  labs(x = "Significance Score")
```

3.1.4 Databases

To confirm the relevance of these TFs, it is necessary to search for information about them in different databases such as *The Human Protein Atlas*, *STRING*, *BioGrid* or *GeneCards* (**Figure 6**).

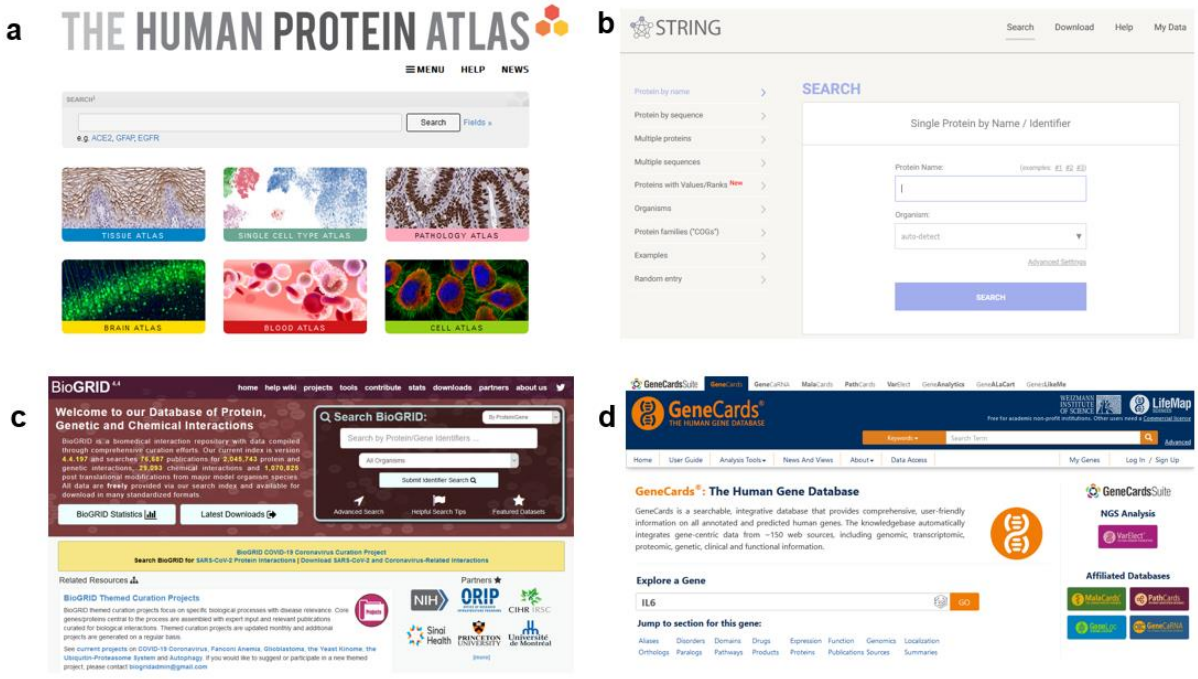

Figure 6. Databases interfaces: (**a**) *The Human Protein Atlas*; (**b**) *STRING*; (**c**) *BioGrid*; (**d**) *GeneCards*.

The Human Protein Atlas (14) is an open access database created to map all the human proteins in cells, tissues and organs using an integration of various omics technologies. It consists in six parts focusing each one on particular aspects of the human proteome, e.g., the Tissue Atlas, which shows information regarding the expression profiles of human genes both on the mRNA and protein level across tissues and organs; or the Single Cell Type Atlas, which contains single cell RNA sequencing (scRNA-seq) data from 13 different human tissues.

STRING (15) is a database of known and predicted protein-protein interactions, including direct (physical) and indirect (functional) associations, from thousands of organisms.

BioGrid (16) and *GeneCards* (17) are two databases which provide information about protein and genetic interactions from major model organism species and information about all annotated and predicted human genes, respectively.

3.2 Identification of TF cell type specific signature in scRNA-seq datasets

A signature score was generated for each cell in single cell data in order to identify the relevance of the TF in a cell specific context. The preparation of single cell RNAseq data and TF signature score was done as described below:

3.2.1 scRNA-seq dataset obtention

The public dataset used in this project was obtained from the next publication: Strunz, Maximilian, et al. "Alveolar regeneration through a Krt8+ transitional stem cell state that persists in human lung fibrosis." Nature communications 11.1 (2020): 1-20.

In this study (18), the group discovered a transcriptional convergence of airway and alveolar stem cells to a Krt8+ transitional stem cell state that precedes the regeneration of AT1 cells. This Krt8+ ADI cell state persists in models of progressive lung fibrosis and human IPF patients.

The dataset of this publication was chosen for our project because it was generated by subjecting sorted cells from lung epithelium to scRNA-seq at 18 time points after injury using two replicate mice per timepoint (**Figure 7**).

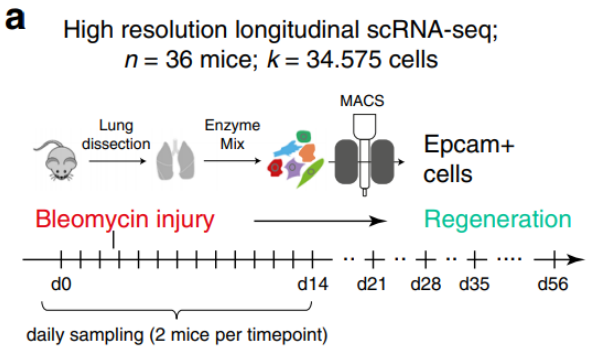

Figure 7. High resolution longitudinal scRNA-seq. Lung injury and pulmonary fibrosis were induced by single-dose administration of bleomycin hydrochloride. Source (18).

3.2.2 scRNA-seq analysis and clustering

The computational analysis of the dataset was largely performed using the *Seurat* package (19) in R Studio. After creating a Seurat object for this sample, cells were filtered based on percent mitochondrial genes in order to exclude damaged and dying cells. The sample was then clustered using the FindClusters() function based on an input number of principal components (PC) and resolution.

Single cell transcriptional profiles were visualized in two dimensions using the Uniform Manifold Approximation and Projection (UMAP) method. Different cell type clusters were annotated according to the steps in the publication and using their interactive webtool (20). To be able to identify cell type clusters prior information is necessary about marker genes of each cell type, which were also obtained from the publication.

3.2.3 Generation of TF signature scores

The TF signature score was generated by evaluation of known target genes of each TF. These gene lists can be found in databases such as *Harmonizome* (**Figure 8**) (21). This database allows to search for information about genes and proteins from over a hundred publicly available resources. The list selected for each TF must contain the target genes associated or the most common protein-protein interactions and they must be manually curated.

The data access can be made through a JSON API format and then read and converted to a gene list using the *rjson* package in R Studio.

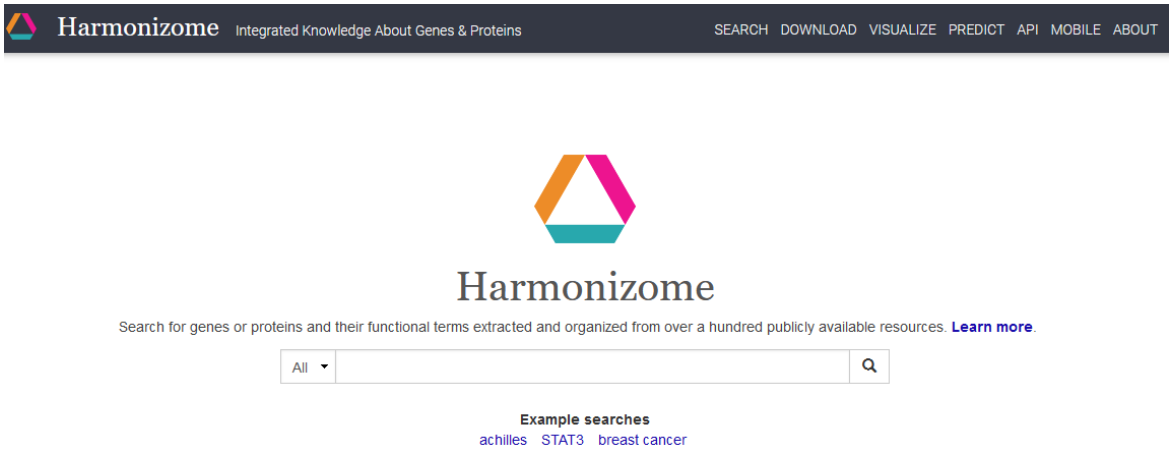

Figure 8. *Harmonizome* database interface.

After collecting all the lists with different gene associations for each TF, a signature score was generated using the function AddModuleScore() in R Studio (22). This function calculates the average expression levels of each cluster on single cell level, subtracted by the aggregated expression of control feature sets. Then this signature score was projected on the single cell UMAP image to identify the most important cell types with a specific TF activity.

4. RESULTS

4.1 Scan analysis and enrichment analysis

As previously explained above, the tracheal epithelial cells datasets from YAP, TAZ and both YAP/TAZ were analysed with the CiiiDER bioinformatic tool. The first analysis developed was the scan analysis, which predicted the potential TFBSs of the different datasets separately. However, some of these predicted TFBSs may be false positives and others may be true in certain biological contexts, so it is important to perform the next enrichment analysis.

The enrichment analysis compares the predicted binding sites to those found in a background list of genes. The results of the enrichment analysis for the YAP/TAZ annotated genes are visualised on this interactive plot (**Figure 9**) with 843 over- and under-represented TFs.

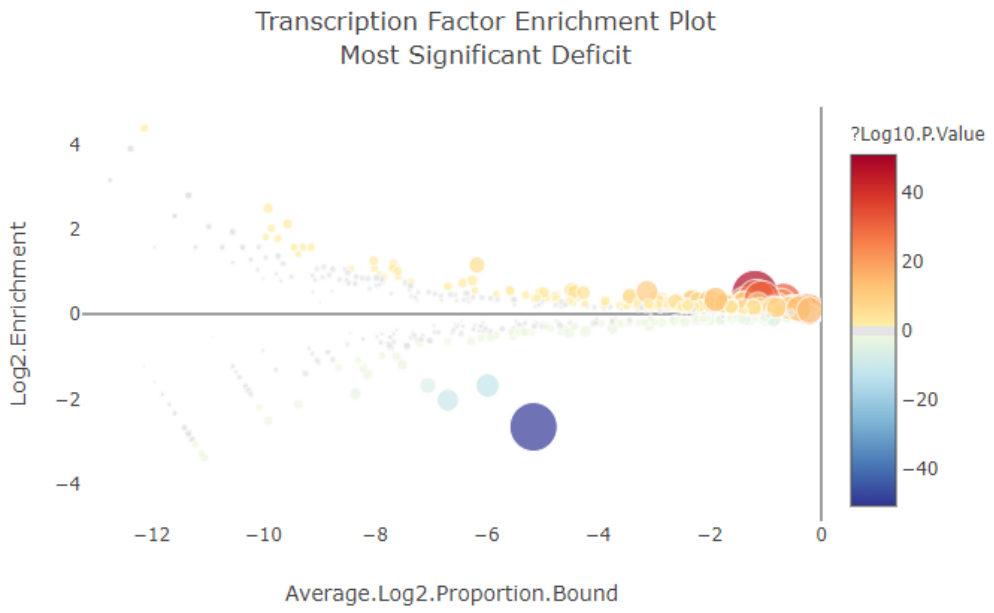

Figure 9. CiiiDER enrichment results for the tracheal cells dataset. The data are derived from the proportion of regions bound for each TF. Size and colours of TFs represent their P-values (Significance Scores).

4.2 Selection of the most relevant TFs

A list of 843 TFs were obtained after the enrichment analysis and then a small group with the most remarkable TFs was selected based on the Significance Scores.

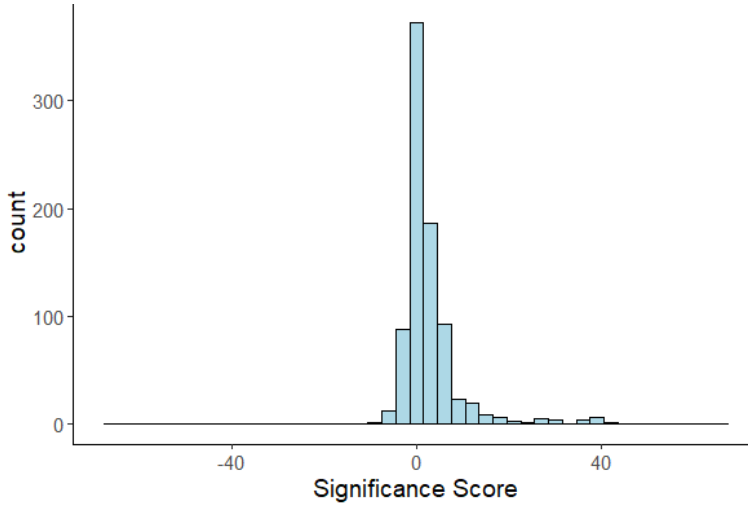

Figure 10. Representation of Significance Scores of 843 TFs.

To do this, we chose the most significant 5% of the data based on the distribution. As the histogram shows (**Figure 10**), the distribution of the Significance Score somewhat follows a gaussian distribution, although is not perfect. We therefore wanted to select a Significance Score that is higher than values under the 95% confidence interval of the distribution, which is calculated based on the mean (μ) and standard deviations (σ): $\mu + 2 \times \sigma = 21.4347$. However, since the distribution is slightly skewed, we also calculated the cut off based on the median: $M_e + 2 \times \sigma = 19.0562$.

Finally, we decided to use the cut off based on the mean (~ 21) as the data is not perfectly gaussian. So as result, a total of 28 TFs were obtained as the most relevant ones (**Figure 11**).

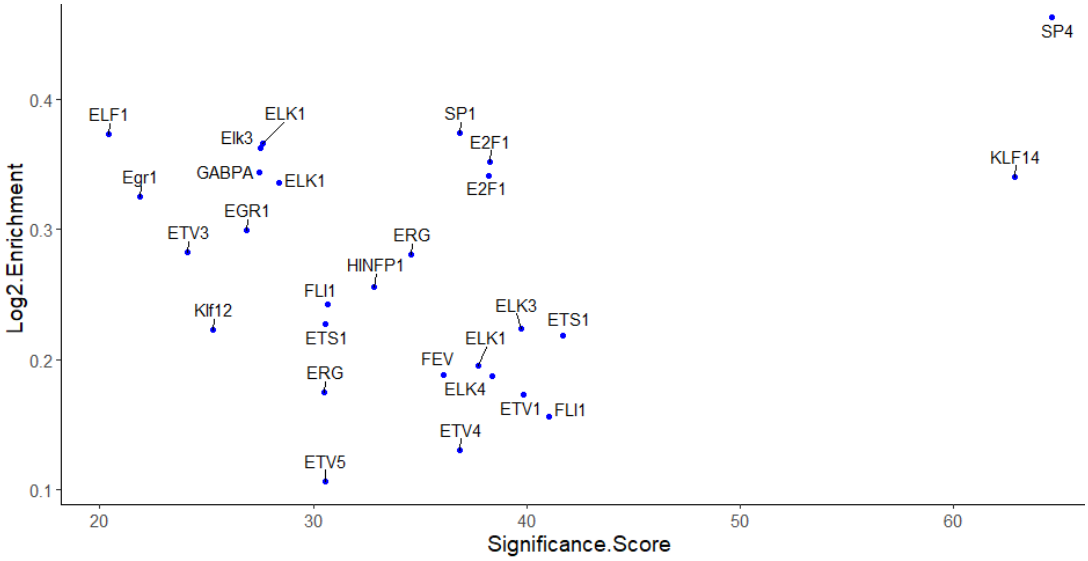

Figure 11. Representation of the selected 28 TFs showing the Significance Score and the Enrichment.

4.3 Exploration of the selected TFs

To explore the 28 TFs, we first analysed their RNA single cell type specificities in *The Human Protein Atlas* database to see whether they are found in a specific cell type such as AT2 cells. We also analysed their protein interactions in *STRING* and *BioGrid* databases to find correlations with YAP and TAZ and searched for information about their cell functions in *GeneCards* database. After this, we concluded that the potentially most interesting TFs found were three among the top 5% of all over-represented TFs: E2F1, ETV1 and FEV.

ETV1 is a member of the ETS (E twenty-six) family of TFs. This family regulates many target genes that modulate biological processes like cell growth, angiogenesis, migration, proliferation and differentiation (17) and this TF is also overexpressed in some cancers. It is considered relevant because its RNA single cell type specificity is AT2 cells according to *The Human Protein Atlas* database (**Figure 12; a**).

E2F1 is a member of the E2F family of TFs, playing a crucial role in the control of cell cycle and action of tumor suppressor proteins, being also a prognostic marker in some cancers (17). It is considered relevant because its RNA single cell type specificity is also AT2 cells according to *The Human Protein Atlas* database (**Figure 12; c**).

FEV belongs to the ETS family of TFs and is a prognostic marker in some cancers (14). The relevance of this TF is its absence in the lung according to *The Human Protein Atlas* database, in contrast to the previous TFs that are found in at least one type of cells of the lung (**Figure 12; e**).

On the other hand, the protein-protein interactions found for each of these 3 TFs in *STRING* database (**Figure 12; b, d, f**) do not show direct or indirect associations to YAP and TAZ, suggesting that these interactions may not have being verified yet.

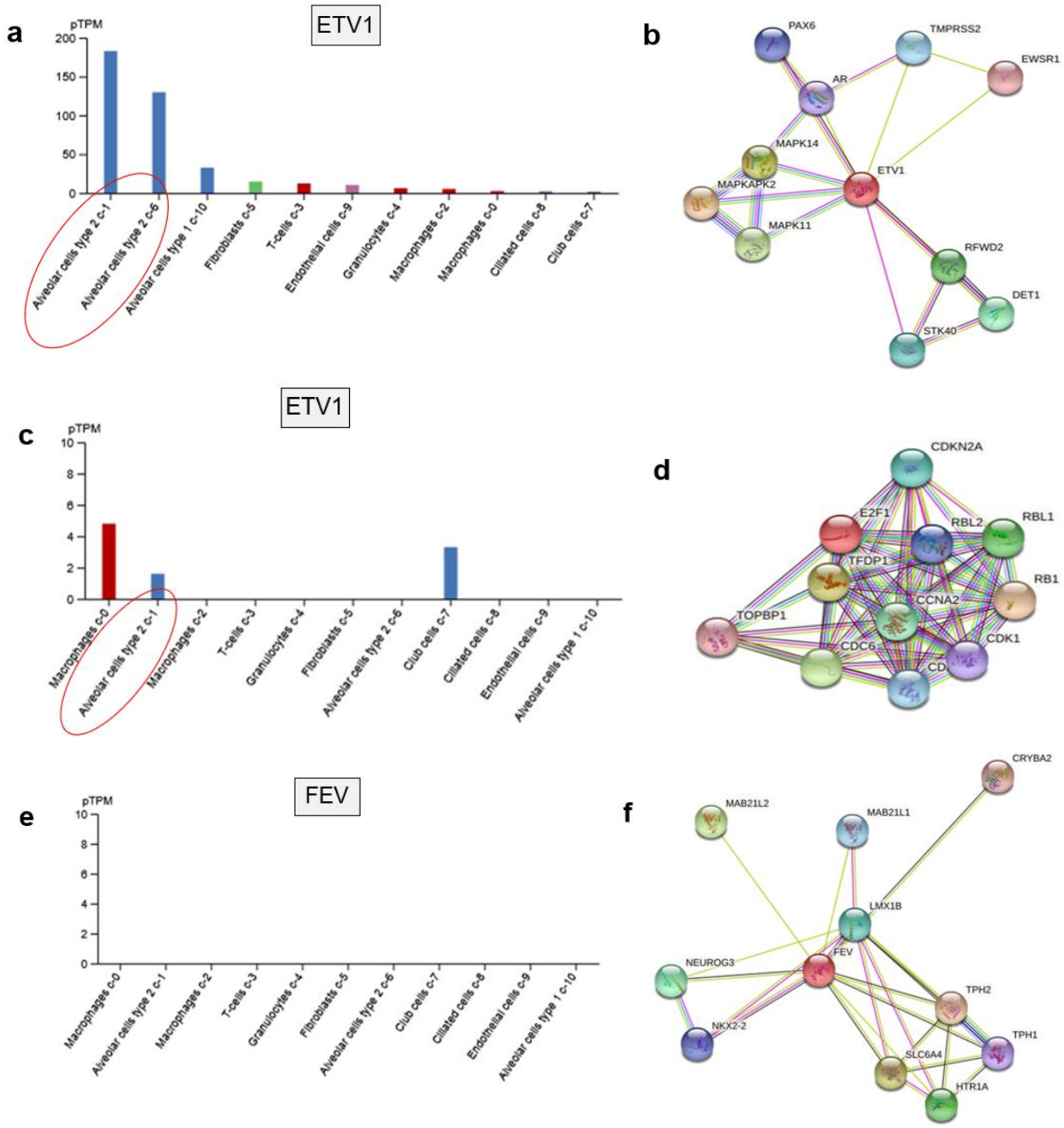

Figure 12. Representation of RNA single cell type specificities in the lung according to *The Human Protein Atlas* database on the left side (14) and representation of protein-protein interactions according to *STRING* database on the right side (15), for ETV1 (**a, b**), E2S1 (**c, d**) and FEV (**e, f**), respectively.

4.4 Identification of TF cell type specific signature in a scRNA-seq dataset

To confirm the expression of the selected TFs in lung epithelial cells, 13 clusters were annotated from a public scRNA-seq dataset using the UMAP method (**Figure 13**).

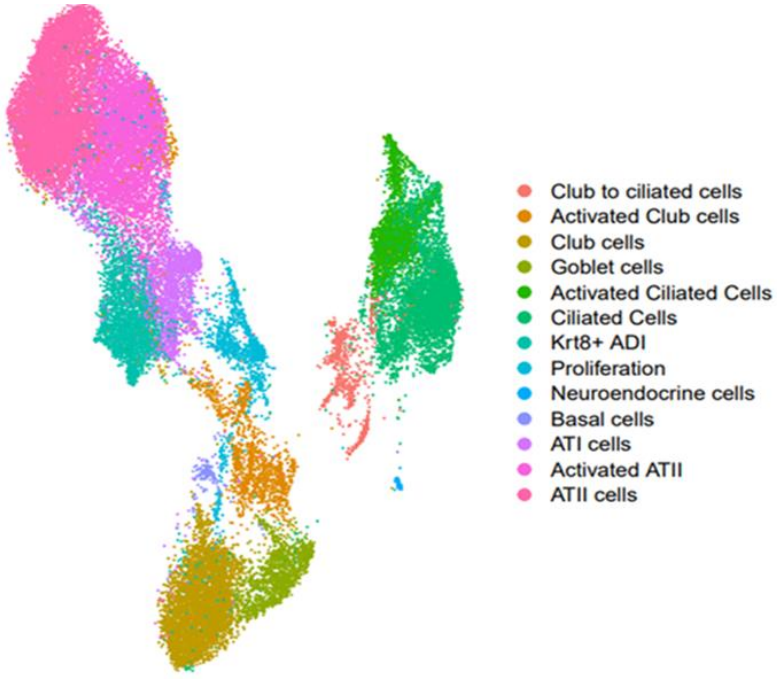

Figure 13. Visualization of the 13 cell type clusters of the lung epithelium annotated. Each colour corresponds to a cell type.

After the clusters were defined, the expression of the most relevant TFs was examined to observe if they are found in the dataset or not. We concluded that these TFs were ETS1, E2F1 and FEV, showing for E2F1 and FEV a low expression in the dataset (**Figure 14; a**). Low RNA expression does not correlate with activity of the transcription factor.

Thus, we aimed at generating a signature score for each TF in order to evaluate its activity. The signature score for these TFs was projected on the single cell UMAP image (**Figure 14, b**). We found ETS1 to have a higher signature score in distal and proximal cells while E2F1 has a higher signature score in proximal and goblet cells. However, FEV's signature score is mainly in proximal epithelial cell types despite its low expression in the scRNA-seq dataset.

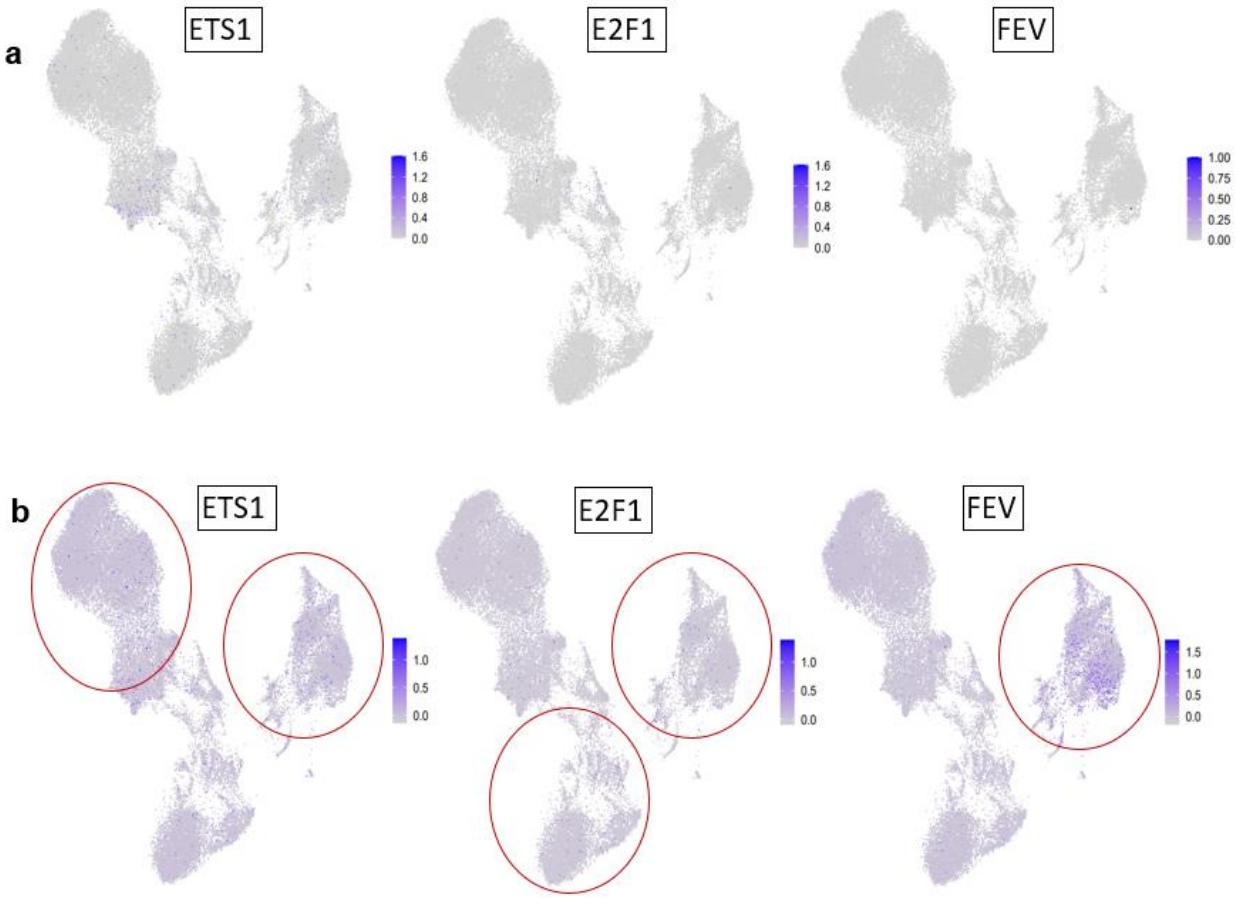

Figure 14. Analysis of a publicly available scRNA-seq dataset using the UMAP method. (**a**) Expression of 3 relevant TFs in the dataset: ETS1, E2F1 and FEV. The expression is low in E2F1 and FEV. (**b**) Projection of the signature score of ETS1, E2F1 and FEV on the single cell UMAP image. ETS1 is expressed in distal and proximal cells. E2F1 is expressed in proximal and goblet cells. However, FEV is expressed mainly in proximal epithelial cell types.

5. DISCUSSION

IPF is a lethal lung disease with increasing incidence and prevalence as several studies worldwide show (2). Currently, the mainstay of pharmacological treatment for IPF is monotherapy with pirfenidone or nintedanib, two antifibrotic drugs approved by the FDA in 2014. Although an improvement over previously suggested therapies, they do not completely arrest or improve lung function and their mechanisms of action in the treatment of IPF remains unclear (23). As there are still no widely effective therapies for this incurable disease, an opportunity for development of novel or add-on pharmacologic agents is presented.

Verteporfin, a benzoporphyrin derivative, is clinically used for the treatment of neovascular macular degeneration. Recent studies described verteporfin as a potent inhibitor of YAP-TEAD complex, which may seem a promising anti-fibrotic strategy. A study by the Wagner lab have shown the anti-fibrotic effects of verteporfin in Mice (unpublished data). Verteporfin has been shown to inhibit interaction with the TEAD family transcription factors (10), however, whether they inhibit other YAP/TAZ interactions is yet unknown. Moreover, the high cytotoxicity of verteporfin makes it less likely to be suitable for use with humans. Additionally, YAP and TAZ interactions are context dependant, making it difficult to target them therapeutically to treat IPF. Therefore, the disadvantages of all-pathway molecular inhibitors make it necessary to develop specific strategies against intracellular and protein-protein processes interactions.

Identification of pathologic TFs interactions causing the disease could potentially help for defining therapeutical targets while maintaining the useful functions of repair-promoting interactions. The optimization of the analysis workflow developed in this thesis has allowed us to identify potential YAP/TAZ binding partners with suitable characteristics to continue with future experimental research in the laboratory to identify which of these interactions are pathologic.

The TFs identified were three: ETS1, FEV and E2F1. The first two belong to the ETS domain family of TFs and the third one belongs to the E2F family (17), being all involved in different oncogenic processes such as breast tumors (24–26) or Ewing tumors (27).

Our results showed that these TFs are active in different cell types of the lung epithelium even though the dataset used for the study was obtained from tracheal cells. Some of the TFs are more active in alveolar cells, others are more active in proximal cells and others in

proximal/goblet cells. This shows that binding interactions between YAP/TAZ are cell type dependent. The activity of TFs can be specific or shared for each cell type depending on the biological context in which they are. Therefore, this may explain that in fibrotic changes of IPF, cells also change their interaction binding partners. This might explain the presences of transitional cells in IPF that are highly transcriptionally active which express several markers of the proximal and distal progenitor cells simultaneously. To confirm this in the future we will do the same analysis for distal (alveolar) cells. This will allow us to compare the interactions between YAP/TAZ and which combinations are active during IPF.

In silico predictive research tools have clearly demonstrated their utility thanks to the simplification of online tools and correlation it provides in handling the large amount of data available nowadays. For this reason, the computational prediction of YAP/TAZ transcription factor binding partners will be an indispensable requirement in research prior to *in vitro* and *in vivo* validation of potential new therapeutic strategies. Afterwards, studies in representative cell and animal models for fibrosis focusing on these YAP/TAZ binding partners are required to completely understand these interactions in lung epithelial cells in the context of fibrosis.

On the other hand, not only YAP/TAZ signaling but several signal transduction pathways have been linked to the pathophysiology of fibrosis, such as transforming growth factor β (TGF-β) or Wingless/Int (WNT). Therefore, the predictive approach for YAP/TAZ developed in this work may provide an efficient computational method that might be replicated for other signalling pathways and different cellular contexts in the future.

6. BIBLIOGRAPHY

1. Gokey JJ, Sridharan A, Xu Y, Green J, Carraro G, Stripp BR, et al. Active epithelial Hippo signaling in idiopathic pulmonary fibrosis. JCI Insight. 2018 Mar 22;3(6).

2. Ley B, Collard HR, King TE. Clinical Course and Prediction of Survival in Idiopathic Pulmonary Fibrosis. Am J Respir Crit Care Med. 2011 Feb 15;183(4):431–40.

3. Parimon T, Yao C, Stripp BR, Noble PW, Chen P. Alveolar Epithelial Type II Cells as Drivers of Lung Fibrosis in Idiopathic Pulmonary Fibrosis. Int J Mol Sci. 21(7).

4. Barkauskas CE, Chung M-I, Fioret B, Gao X, Katsura H, Hogan BLM. Lung organoids: current uses and future promise. Development. 2017 Mar 15;144(6):986–97.

5. Barkauskas CE, Noble PW. Cellular Mechanisms of Tissue Fibrosis. 7. New insights into the cellular mechanisms of pulmonary fibrosis. Am J Physiol-Cell Physiol. 2014 Apr 16;306(11):C987–96.

6. Barkauskas CE, Cronce MJ, Rackley CR, Bowie EJ, Keene DR, Stripp BR, et al. Type 2 alveolar cells are stem cells in adult lung. J Clin Invest. 2013 Jul 1;123(7):3025–36.

7. Plosa E, Guttentag SH. 42 - Lung Development. In: Gleason CA, Juul SE, editors. Avery's Diseases of the Newborn (Tenth Edition). Philadelphia: Elsevier; 2018. p. 586-599.e2.

8. Xu Y, Mizuno T, Sridharan A, Du Y, Guo M, Tang J, et al. Single-cell RNA sequencing identifies diverse roles of epithelial cells in idiopathic pulmonary fibrosis. JCI Insight. 2017 Mar 16;1(20).

9. Dey A, Varelas X, Guan K-L. Targeting the Hippo pathway in cancer, fibrosis, wound healing and regenerative medicine. Nat Rev Drug Discov. 2020 Jul;19(7):480–94.

10. Piersma B, Bank RA, Boersema M. Signaling in Fibrosis: TGF-β, WNT, and YAP/TAZ Converge. Front Med. 2015;2.

11. Skene PJ, Henikoff JG, Henikoff S. Targeted in situ genome-wide profiling with high efficiency for low cell numbers. Nat Protoc. 2018 May;13(5):1006–19.

12. CiiiDER: A tool for predicting and analysing transcription factor binding sites [Internet]. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0215495

13. CiiiDER [Internet]. Available from: http://www.ciiider.org/

14. The Human Protein Atlas [Internet]. Available from: https://www.proteinatlas.org/

15. STRING: functional protein association networks [Internet]. Available from: https://string-db.org/cgi/input?sessionId=bFWk5VuTKbYf&input_page_show_search=on

16. BioGRID | Database of Protein, Chemical, and Genetic Interactions [Internet]. Available from: https://thebiogrid.org/

17. GeneCards - Human Genes | Gene Database | Gene Search [Internet]. Available from: https://www.genecards.org/

18. Strunz M, Simon LM, Ansari M, Kathiriya JJ, Angelidis I, Mayr CH, et al. Alveolar regeneration through a Krt8+ transitional stem cell state that persists in human lung fibrosis. Nat Commun. 2020 Jul 16;11(1):3559.

19. Tools for Single Cell Genomics [Internet]. Available from: https://satijalab.org/seurat/

20. Mouse Lung Injury & Regeneration – Schiller & Theis labs @ Helmholtz Center Munich [Internet]. Available from: http://146.107.176.18:3838/Bleo_webtool_v2/

21. Harmonizome [Internet]. Available from: https://maayanlab.cloud/Harmonizome/

22. AddModuleScore function - RDocumentation [Internet]. Available from: https://www.rdocumentation.org/packages/Seurat/versions/4.0.2/topics/AddModuleScore

23. Saito S, Alkhatib A, Kolls JK, Kondoh Y, Lasky JA. Pharmacotherapy and adjunctive treatment for idiopathic pulmonary fibrosis (IPF). J Thorac Dis. 2019 Sep;11(Suppl 14):S1740–54.

24. Sun C-C, Li S-J, Hu W, Zhang J, Zhou Q, Liu C, et al. Comprehensive Analysis of the Expression and Prognosis for E2Fs in Human Breast Cancer. Mol Ther J Am Soc Gene Ther. 2019 Jun 5;27(6):1153–65.

25. Sinh ND, Endo K, Miyazawa K, Saitoh M. Ets1 and ESE1 reciprocally regulate expression of ZEB1/ZEB2, dependent on ERK1/2 activity, in breast cancer cells. Cancer Sci. 2017;108(5):952–60.

26. Dittmer J. The role of the transcription factor Ets1 in carcinoma. Semin Cancer Biol. 2015 Dec;35:20–38.

27. Peter M, Couturier J, Pacquement H, Michon J, Thomas G, Magdelenat H, et al. A new member of the ETS family fused to EWS in Ewing tumors. Oncogene. 1997 Mar;14(10):1159–64.

7. ANNEXES

Table with the selected 28 transcriptions factors and the information obtained in the enrichment analysis.

Transcription Factor	Significance Score	Average Proportion Bound	Enrichment
SP4	64,62096	-1,07692	0,463211
KLF14	62,85525	-0,68951	0,340636
ETS1	41,68062	-0,4642	0,218372
FLI1	41,04773	-0,25941	0,156387
ETV1	39,86491	-0,322	0,173472
ELK3	39,75714	-0,50642	0,224367
ELK4	38,37832	-0,3816	0,187232
E2F1	38,24728	-1,06086	0,351542
E2F1	38,21647	-1,0189	0,341557
ELK1	37,7257	-0,41758	0,195279
SP1	36,86953	-1,1877	0,37384
ETV4	36,83855	-0,20577	0,130689
FEV	36,10375	-0,40617	0,187994
ERG	34,59159	-0,81916	0,280743
HINFP1	32,84704	-0,73839	0,255717
FLI1	30,64113	-0,72114	0,24305
ETS1	30,53861	-0,64963	0,227615
ETV5	30,53347	-0,16789	0,106726
ERG	30,50237	-0,41681	0,174982
ELK1	28,388	-1,23265	0,336129
ELK1	27,63339	-1,40606	0,366094
Elk3	27,51674	-1,39128	0,362619
GABPA	27,4732	-1,30058	0,34394
EGR1	26,86713	-1,09056	0,299373
Klf12	25,31188	-0,73395	0,222937
ETV3	24,08341	-1,08683	0,282304
Egr1	21,89246	-1,40909	0,325171
ELF1	20,39536	-1,74306	0,373014