Epithelial Carcinogenesis Group

Network analysis of transcriptional regulation of acinar cell identity using transcription factor footprinting inference from ATAC-seq data

Máster en Bioinformática y Bioestadística

Autor:
Francisco Javier Soriano Díaz

Tutores:
Jaime Martínez de Villarreal
Laia Bassaganyas Bars

Junio de 2022


BY NC SA

FINAL WORK CARD

Title:	Network analysis of transcriptional regulation of acinar cell identity using transcription factor footprinting inference from ATAC-seq data
Author:	Francisco Javier Soriano Díaz
Tutor:	Jaime Martínez de Villarreal Laia Bassaganyas Bars
Date of delivery:	Junio de 2022
Studies:	Máster en Bioinformática y Bioestadística
Language:	English
Number of credits:	15
Keywords:	transcriptional regulation, network, acinar identity

Abstract

La identidad celular puede considerarse un importante mecanismo supresor de tumores. En este contexto, el esclarecimiento de los mecanismos reguladores de dicha identidad, fundamentalmente de la actividad de los factores de transcripción que regulan la expresión de distintos programas transcripcionales, resulta esencial. Tomando como punto de partida la estructura tridimensional de la cromatina a partir de los datos proporcionados por la técnica *Assay for Transposase-Accessible Chromatin using sequencing* (ATAC-seq), se pretenden esclarecer las redes transcripcionales implicadas en el mantenimiento de las condiciones homeostáticas en páncreas murino para posteriormente reconocer aquellos programas transcripcionales implicados en la pérdida de identidad celular necesaria en el proceso carcinogénico. Para ello, se dispone de datos provenientes de animales modificados genéticamente usados como modelos en el estudio del adenocarcinoma pancreático ductal (*pancreatic ductal adenocarcinoma*, PDAC). Con este trabajo se espera tener una mejor comprensión de las redes de regulación génica y por tanto de la relación existente entre los factores de transcripción, sus lugares de unión y los genes involucrados en el cáncer estudiado. Con la creación de estas redes se podrán confirmar resultados obtenidos experimentalmente, así como servir de base para nuevas investigaciones, estableciéndose una relación bidireccional entre el trabajo computacional y el realizado en el laboratorio. Asimismo, se espera poder presentar la información aquí obtenida como un recurso que pueda ser empleado por otros investigadores para sus trabajos. Todo ello tiene el objetivo final de conocer mejor y combatir el cáncer de páncreas.

Abstract

Cell identity can be considered an important tumor suppressor mechanism. In this context, the clarification of the regulatory mechanisms of said identity, fundamentally of the activity of the transcription factors that regulate the expression of different transcriptional programs, is essential. Taking the three-dimensional structure of the chromatin as a starting point from the data provided by the technique Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq), the aim is to clarify the transcriptional networks involved in the maintenance of homeostatic conditions in the murine pancreas to later recognize those transcriptional programs involved in the loss of necessary cellular identity in the carcinogenic process. For this, data are available from genetically modified animals used as models in the study of pancreatic ductal adenocarcinoma (PDAC). With this work, it is expected to have a better understanding of gene regulation networks and therefore of the relationship between transcription factors, their binding sites and the genes involved in the studied cancer. With the creation of these networks, results obtained experimentally can be confirmed as well as serve as a basis for new research, establishing a bidirectional relationship between computational work and that carried out in the laboratory. Likewise, it is expected that the information obtained can be presented as a resource that can be used by other researchers for their work. All of this has the ultimate goal of better understanding and combating pancreatic cancer.

Acknowledgments

First of all, I would like to express my deep gratitude to Jaime Martínez de Villarreal who has been my guide during these months. His help and his knowledge have been vital to the accomplishment of this work.

The person who made it possible for me to be part of this magnificent experience in a center of excellence such as the CNIO was Francisco X. Real. He was the one who welcomed me and guided me to enter there so my following words of thanks go to him.

I also have to mention Mónica, since it was she who generated much of the experimental data and without which this work would not have been the same. Thank you very much for your kindness and willingness to help.

Of course, I cannot forget to mention also the rest of the CNIO Epithelial Carcinogenesis group I have been part of during this time. I felt integrated from day one and have really enjoyed his company.

And I cannot fail to mention the Bioinformatics Unit of the CNIO, the colleagues with whom I have shared the day to day, and specially Manu, for his joy and good humor, and Ester, who I am sure will do a wonderful job and who has been lovely from day one.

I have not had the pleasure of meeting Pablo Perez in person, but his results and advice at the beginning have been a determining part of this work. Thank you very much.

I also wanted to thank the members of the UOC, specially Laia Bassaganyas Bars for her interest and kindness in helping me when I needed it.

Finally, I have to thank my family and friends who have been as enthusiastic about my work as I have been during these months.

Contents

1 Introduction **12**

2 Results **15**

2.1 Homeostatic pancreas 15

2.2 Perturbation data 19

2.3 Intersection of the networks 21

2.4 Topological study of subnetworks 22

2.5 Individual TF subnetworks 23

2.6 Web application 26

3 Discussion **31**

4 Methods **34**

4.1 Datasets 34

4.2 ATAC-seq analysis 34

4.3 Footprinting analysis 35

4.4 Peak merging and annotation 35

4.5 Transcription factor binding motifs 36

4.6 Network visualization and analysis 36

4.7 Web application development 37
4.8 Code availability 37

List of Figures

1 Data analysis of homeostatic conditions 18

2 Venn diagrams displaying pairwise comparisons of perturbation specific networks 22

3 Venn diagrams for all pairwise comparison where the edges are compared 24

4 Study of transcriptional networks 25

5 ATAC-seq footprint network analyzer 29

6 ATAC-seq footprint network analyzer. Functional analysis 30

1 *Supplementary figure.* Motif analysis of Ptf1a 44

2 *Supplementary figure.* Density plot of gene expression in acinar cells and table with first top 100 expressed TFs 45

3 *Supplementary figure.* Variation of the position occupied by the TFs in the lists ranked by their degree for the pairwise comparison between networks 46

List of Tables

1 Nodes and edges for homeostatic networks and all perturbation data networks 20

1 *Supplementary table.* The 299 TFs selected, from TFs expressed in acinar cells, as input for the building of transcription factor networks 46

Abbreviations

- **ADM:** Acinar to Ductal Metaplasia
- **ATAC-seq:** Assay for Transposase-Accessible Chromatin using sequencing
- **BAM:** Binary Alignment Map
- **CAE:** Caerulein
- **ChIP-seq:** Chromatin Immunoprecipitation-coupled with Sequencing
- **DWM:** Dinucleotide Weight Matrix
- **FACS:** Fluorescence Activated Cell Sorter
- **GEMM:** Genetically Engineered Mouse Model
- **GRN:** Gene Regulatory Network
- **IDR:** Irreproducible Discovery Rate
- **NGS:** Next-Generation Sequencing
- **NT:** No Treatment
- **OCR:** Open Chromatin Region
- **PanIN:** Pancreatic Intrapithelial Neoplasia
- **PBS:** Phosphate Buffered Saline
- **PDAC:** Pancreatic Ductal Adenocarcinoma
- **PWM:** Position Weight Matrix
- **scRNA-seq:** single cell RNA sequencing
- **TF:** Transcription Factor

- **TG:** Target Gene
- **TFBS:** Transcription Factor Binding Site
- **TSS:** Transcription Start Site
- **TTS:** Transcription Termination Site
- **WT:** Wild Type

1. Introduction

The pancreas is a glandular organ present in vertebrates that is divided into exocrine and endocrine functional components. Only 5% of its mass is made up of endocrine cells that form structures called islets of Langerhans which produce and secrete insulin and glucagon that regulate glucose homeostasis. Therefore, most of this organ is made up of exocrine cells that synthesize the hydrolytic digestive enzymes that are transported to the intestine where they contribute to the digestion of carbohydrates, proteins and lipids. Exocrine cells are classified into acinar and ductal cells, the former being specialized in synthesizing, storing and secreting digestive enzymes while the latter form the ducts that transport them to the duodenum [1] [2].

Pancreatic ductal adenocarcinoma (PDAC) represents 90% of cancers arising in this organ. It is highly aggressive, with a mean survival time of 5 months following diagnosis and a 5-year survival of 5%, due to a lack of early diagnosis and poor response to treatments [3] [4].

The phenotype of the cells that constitute PDAC is generally ductal, both in terms of morphology and antigen expression. There is evidence in genetically engineered mouse models (GEMMs) that PDAC can originate from all exocrine cells [5] [6] [7] [8]. Its origin is debated although it has been described that it may be a consequence of the loss of cellular identity of the acinar cells by a process known as acinar to ductal metaplasia (ADM). As a consequence of this process, acinar cells transdifferentiate into ductal-type cells. Oncogenic genetic insults and environmental stress can promote ADM to pancreatic intraepithelial neoplasia (PanIN) [9], a precancerous lesion.

Loss of acinar identity is considered a starting point of carcinogenesis, as a consequence of the tissue damage to which it is associated [10] [11]. This identity is controlled by specific gene expression which, in turn, is regulated by the interaction of transcription factors (TFs) [12], proteins that bind to specific DNA sequences controlling the transcription of genetic information from DNA to RNA. These TFs

bind to DNA in cis regulatory elements such as enhancers and promoters and, according to the TF, they up-regulate or down-regulate the gene whose expression they are controlling. Acinar identity is driven by specific genetic programs that group several genes controlled by well-defined DNA-binding TFs, and these in turn can be classified into transcriptional modules, a set of genes co-regulated by a single TF. The ability to remain in their differentiated state of acinar cells has been suggested to act as a suppressor mechanism for tumor processes [13] [14], which implies that knowing the transcriptional modules associated with this process is essential to better understand the mechanisms that give rise to cancer.

Biological interactions associated with regulatory mechanisms are highly complex and therefore their study is not trivial. However, thanks to the computational advances associated with biology, strategies such as Gene Regulatory Networks (GRNs) [15] [16] [17] [18] have been developed to help with this purpose, which from experimental data allow inferring biological behaviors and functions. The use of networks together with the use of Next-Generation Sequencing (NGS) technologies allows the study and understanding of biological mechanisms such as transcriptional regulation, modeling their behavior and the interactions between the different actors that are part of it.

In order to study transcriptional regulation, GEMMs that mimic the initial steps of carcinogenesis can be used. These models may carry mutations present in human PDAC such as the G12V activating mutation of *Kras* [19] oncogene, or knockout of genes, such as *Gata4* and *Gata6*, involved in epithelial differentiation in the pancreas [20] [21].

One way of approaching the study of transcriptional regulation is by using ATAC-seq [22] and footprinting. ATAC-seq is an experimental procedure by which the accessibility of open chromatin for the entire genome is studied.

This work addresses a study of the transcriptional regulation of the mouse pancreas when it is subjected to perturbations (mutated *Kras*, knockout of *Gata4*, knockout of *Gata6* and induced pancreatitis) in order to analyze acinar identity. For this, NGS technologies and bioinformatics tools have been used to generate transcriptional networks from the data. From ATAC-seq dataset, the open chromatin regions (OCRs) [23] are studied, which allow to perform a footprinting analysis and finally to locate the transcription factor binding sites (TFBSs). From this information, the relationship between TF and genes can be inferred and transcriptional networks can be generated. With this approach, the regulation of acinar identity has been studied under homeostatic conditions and in contexts in which it is challenged by specific genetic insults, by the inflammation of the pancreas (pancreatitis), or by both possibilities.

The networks generated in this work can be used as a source of information to define transcriptional modules and their biological functions for various perturbation situations that are precursors of PDAC. It can also serve as a bidirectional tool to both validate and generate hypotheses.

2. Results

2.1 Homeostatic pancreas

ATAC-seq is an experimental procedure by which the accessibility of open chromatin for the entire genome is studied. OCRs are transcriptionally active areas of DNA as they allow access to RNA polymerase and allow the process of DNA transcription to begin. The hyperactive Tn5 transposase cuts and inserts adapters into regions where chromatin is accessible. The OCRs are defined from the insertion signal of Tn5 since it is here where the TFs bind to the genome. A decrease in signal indicates the presence of a portion of DNA bound to proteins and therefore Tn5 cannot cut this region. These areas where the signal decreases are called footprints [24] and will be used to determine how TFs interact with the genome.

The analysis of the pancreatic homeostatic data was performed to test the methodology and to check the stability of the data. For this purpose, a dataset previously studied in another study was analyzed for cross-validation. The data were obtained from a publicly accessible ATAC-seq atlas [25] consisting of 66 profiles from 20 different tissues. The four replicates corresponding to the pancreas were used, two of them from male mice and the other two from female mice.

The raw data was processed using the ENCODE ATAC-seq pipeline developed by Anshul Kundaje's laboratory [26]. This pipeline performs the alignment of short-read sequencing data contained in raw FASTQ files to obtain the Binary Alignment Map (BAM) and Browser Extensible Data (BED) files. The BAM files contain the mapped reads to the reference genome and from them the footprint analysis of the accessible regions is carried out to obtain the BED files which hold the coordinates of the regions of enrichment or peaks. These peaks files were extracted for female and male and the consistency of the peak calls between replicates was ensured taking as threshold 0.05 of Irreproducible Discovery Rate (IDR) [27]. To ensure that the signals were biologically relevant and did not contain erroneous data due to noise, a merge was performed between the male replicates, on one side, and the female

replicates, on the other side, using the mergePeaks function of the HOMER bioinformatics software [28]. To carry out these merges, it was taken into account that that there was a large overlap between the two replicates of each sex that allowed the merge to be performed with the guarantee that no information would be lost.

In order to increase the information used with respect to the previous analysis, in which only the OCRs from the intersection of the two signals were considered, it was decided to take into account all existing OCRs in the signals. For this purpose, the intersection between the two signals obtained previously, one for male and the other for female, was calculated and two sets of data were extracted. One corresponding to the intersection of both datasets plus the exclusive data of male and the other to the intersection plus the exclusive data of female. This procedure resulted in 56,249 OCRs for the male and intersection data and 42,504 OCRs for the female and the intersection, compared to the 38,424 OCRs obtained in the analysis prior to this work. Therefore, the number of OCRs increased by 46.39% for the male data and 10.62% for the female data. Therefore, the starting data available were four BAM files (two for male and two for female), which were not modified in any way, and two BED files with the peaks (one for male and one for female), obtained after performing the merges.

To find the TFBSs in the detected OCRs, a footprinting analysis was performed by studying the Tn5 cut signal in order to locate the areas where there was signal depletion in the accessible regions, which would indicate protein binding to DNA. To carry out this task, TOBIAS [29] was used, a bioinformatics toolkit specifically designed to perform a footprinting analysis from the ATAC-seq signal. The first step of the analysis consisted in correcting the ATAC-seq signal since the Tn5 transposase prefers for specific areas of the DNA [30] [31]. This causes a sequence-dependent transposition site bias that distorts the input information and alters the identification of the footprints [32] [33]. To correct this bias the TOBIAS ATACorrect module was used, which takes as input arguments the ATAC-seq reads, the peak files of the areas of interest and a dinucleotide weight matrix (DWM) [34] to generate the expected Tn5 insertion signal for each region. This signal is subtracted from the input data obtaining the corrected signal (Figure 1A). To determine which regions of the signal obtained were footprints, the TOBIAS ScoreBigwig tool was used to evaluate them and obtain the footprint score for each of them. The result is obtained by calculating the difference between the background mean signal and the footprint mean signal, which considers the flanking regions that help to locate footprints whose signal is not so clear (Figure 1B).

The scores obtained were then associated with the TF binding motif data to calculate the specific binding coordinates of each TF. However, before doing this, the list of TFs was restricted to study only those present in acinar cells with the

aim of being more precise in the analysis of the transcriptional regulation involved in cell differentiation. The TFs involved in the acinar activity of the pancreas were chosen from the analysis of RNA-seq and scRNA-seq datasets. These datasets were obtained in our own laboratory in experiments performed on the pancreas of wild type (WT) mice. From the signal of the RNA-seq, the TFs present in WT mouse pancreas were ranked according to their expression levels. Subsequently, a fine adjustment was carried out with the scRNA-seq data [35] allowing to choose only those TFs present in acinar cells. A threshold of 3 RPKM was established to differentiate between true biological signals from noise (Supplementary Figure 2). The complete list of selected TFs can be consulted in Supplementary Table 1.

Once the list of TFs was obtained, they were associated with their corresponding Position Weight Matrix (PWM). The PWMs are matrices which contain the logarithmic probability of the presence of each nucleotide in every one of the positions of a particular motif (Figure 1C). They were obtained from CIS-BP [36] and JASPAR [37]. In order to associate the motifs with the footprints and integrate these different sources of information to predict the TFBSs, the TOBIAS BINDetect tool was used. The result provided by it allows discriminating between bound and unbound TFBSs by setting a threshold value.

For the construction of the networks, it was necessary to define the pair formed by each TFBS and their target gene (TG). This task was carried out with the HOMER annotation function to associate peaks with the closest gene. Once these relationships were obtained, networks were modeled with the TOBIAS CreateNetwork tool. Four TF-TG networks were generated from the obtained data, two for the male replicates and another two for those of the female. These networks are defined by the *nodes*, which represent the regulated or regulator genes, according to whether they are regulated by any of the TFs studied or not. When they are incoming nodes, that is, when the connections leave them towards the gene of interest, they must be interpreted as regulator genes and when they are outgoing nodes, that is, when they receive the connections, they act as regulated genes. These connections between nodes are called *edges* and represent the regulation of one gene, which codes for a TF, on another gene. The number of regulatory processes between genes in the transcriptional networks was recorded, as this value was used to check how the conformation of the networks changed when the input data were modified.

At this point, two networks were available, the network corresponding to the previous work, in which only the OCRs contained in the intersection between the male and female peaks were taken into account, and the one obtained for the work developed so far in which all the OCRs present at the union between the peaks were used. The number of genes (nodes) and regulatory relationships (edges) of both networks was very similar, with a 6.97% increase for the first parameter (10,701 vs.

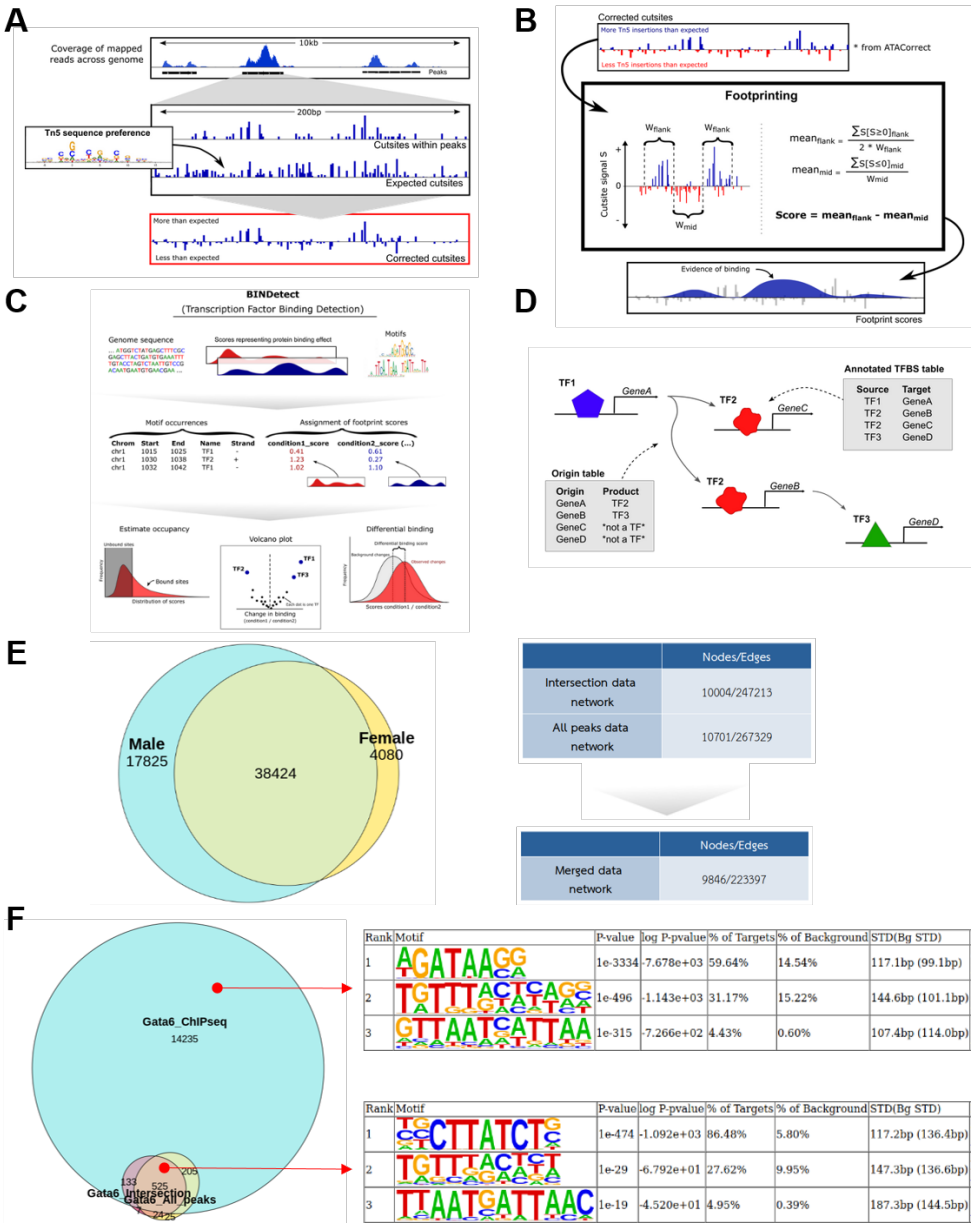


Figure 1: Data analysis of homeostatic conditions. A) Correction of the ATAC-seq signal because of Tn5 bias. **B)** Score of footprints obtaining the difference between the mean background signal and the mean signal of the footprint. **C)** BINDetect module. Association between footprints scores and PFMs. **D)** CreateNetwork module. TFBS-target gene data pairs. **E)** Venn diagram with OCRs for male and female replicates. Nodes and edges for intersection network, all peaks network and the merge of the both previous networks. The network resulting from the merge is very similar in size to the two previous networks, which shows the robustness of the network with all peaks despite having used more OCRs than the intersection network. **F)** Venn diagram of the TFBS coordinates for the intersection, all peaks, and ChIP-seq data. ChIP-seq motif analysis and overlap. The first motif enriched in the overlap coincides with 86.48% of the targets for the TF studied, which shows that the ATAC-seq data are reliable. Figures 1A,1B, 1C and 1D have been modified from Bentsen, M. *et al.*, 2020.

10,004) and an 8.14% increase for the second (267,329 vs. 247,213). In order to check whether the results obtained in the second network were similar with respect to the results of the first network from a qualitative point of view, an intersection was performed between the two networks. The number of genes and of regulatory relationships did not change significantly with respect to the two original networks. The first parameter was reduced by 1.58% (9,846 vs. 10,004) with respect to the first network and by 7.99% (9,846 vs. 10,701) with respect to the second one; the second parameter was reduced by 9.63% (223,397 vs. 247,213) with respect to the first network and by 16.43% (223,397 vs. 267,329) with respect to the second one (Figure 1E).

Once it was confirmed that the process followed to obtain the network provided reliable results, a further analysis was performed to ensure that the data obtained by inference were supported by biological data. In order to cross-validate the footprint strategy, ChIP-seq data from different TFs with a relevant role in pancreatic differentiation acinar identity were taken advantage of. As shown in Figure 1F, the vast majority of GATA6 footprints found in ATAC-seq signal were included in the corresponding ChIP-seq peaks. Additionally, motif analysis from these compartments confirmed the high specificity of the footprinting approach as demonstrated with the higher percentage of TF GATA6 motif found in OCRs (86.48%) versus TF GATA6 ChIP-seq peaks (59.64%). The strategy followed in this work to define the interactions between TFs may be less sensitive and have less resolution than an experimental analysis such as ChIP-seq, but according to the data obtained, it is nevertheless more specific in its results.

Therefore, this first analysis has shown that the procedure followed allows the construction of robust transcriptional networks from the data of an ATAC-seq analysis. Comparison with experimental data ensures that these data are reliable and can be used to study the behavior of acinar identity in the pancreas.

2.2 Perturbation data

Once our strategy was defined and validated under homeostatic conditions, the next step of this research was to determine how transcriptional regulation was affected when acinar identity was disturbed. A set of seven datasets generated within our own research group was studied. These seven datasets simulate perturbation situations that are precursors in the appearance of PDAC: mutated *Kras*, knockout of *Gata4*, knockout of *Gata6* and pancreatitis. Unlike the data on homeostatic conditions which came from a disaggregation of the complete pancreatic tissue, in this case the Fluorescent Activated Cell Sorter (FACS) technique [38] has been used

to differentiate between the different cell types of the pancreas and obtain only the data corresponding to the epithelial cells.

Specifically, the datasets used are NT p48Cre (2 replicates), NT p48Cre; NT *Gata4*KO (3 replicates), NT p48Cre;NT*Gata6*KO (2 replicates), PBS *Kras** (2 replicates), PBS *Kras**;NT*Gata4*KO (3 replicates), Cae *Kras** (3 replicates) and Cae *Kras**;NT*Gata4*KO (3 replicates). P48Cre indicates that a Cre recombinase cDNA has been inserted into the first coding exon of the *Ptf1a* gene which it is useful to induce pancreas-specific recombination [39]. *Kras** represents mutated *Kras* with the mutation G12V [19] [40].

As with the homeostatic data, the same procedure was followed to generate the networks for the perturbation data. After obtaining the footprints, obtained from ATAC-seq analysis, their coordinates were annotated in order to assign each region to the gene that is closest to it. The motifs present in each footprint were determined by their PWM, thus allowing transcriptional regulations to be defined. This process was performed for each of the dataset replicates, so a network was generated for each of them. In order to obtain a single network per dataset, the intersection between each of its replicates was carried out.

Table 1 shows the results obtained in terms of the number of genes and transcriptomic regulations in the networks obtained from the perturbation data and from homeostatic data.

	Initial number of OCRs	Nodes/Edges TF-TG
Homeostatic	60329	10662/247377
NT_p48Cre	78043	12538/421127
NT_*Gata6*KO	68042	13438/434027
NT_*Gata4*KO	82060	13648/554632
PBS_*Kras**	72163	12933/454601
PBS_*Kras*_Gata4*KO	85412	13770/600016
Cae_*Kras**	83334	16214/870194
Cae_*Kras*_Gata4*KO	122822	15872/869532

Table 1: Nodes and edges for homeostatic networks and all perturbation data networks.

The first entry in the table, which refers to the network generated under homeostatic conditions, shows the lowest number of OCRs, 60,329, while the situation with the most perturbations, that is, the dataset in a context of mutated *Kras*, *Gata4*KO and pancreatitis, the highest number, 122,822. According to the results obtained, it was observed that as the number of perturbations increases, so does the number of regulatory relationships between genes of each network. For example, comparing NT_*Gata4*KO (one perturbation) and PBS_*Kras**_*Gata4*KO (two

perturbations) there was a 4.1% increase in the number of OCRs (85,412 versus 82,060). The same situation occurred when comparing PBS_*Kras**_*Gata4*KO (two perturbations) with Cae_*Kras**_*Gata4*KO (three perturbations), with a remarkable 43.8% increase (122,822 versus 85,412).

As a consequence of the increase in OCRs, the topological characteristics of the networks also changed. Using the same example datasets, it was observed that among the datasets NT_*Gata4*KO (one perturbation) and PBS_*Kras**_*Gata4*KO (two perturbations) there was a 0.89% increase in the number of genes (13,770 versus 13,648) and 8.18% increase in the regulatory relationships between genes (600,016 versus 554,632). The same situation occurs when going from PBS_*Kras**_*Gata4*KO (two perturbations) to Cae_*Kras**_*Gata4*KO (three perturbations), with a 15.27% increase in the number of genes (15,872 versus 13,770) and a 44.92% increase in the regulatory relationships between genes (869,532 versus 600,016).

This suggests that as the pancreas was challenged with more perturbations, the transcriptomic stability of acinar identity was reduce and new transcriptional programs were activated, thus generating a higher level of network complexity.

2.3 Intersection of the networks

In order to study the behavior of the pancreas under the action of different perturbations, pairwise comparisons were made between the networks. Each of the comparisons performed was aimed at isolating the effect of the perturbations studied.

The following comparisons were studied: for situations with one perturbation NT p48Cre;NT*Gata4*KO vs NT p48Cre (*Gata4* deletion specific networks), NT p48Cre;NT*Gata6*KO vs NT p48Cre (*Gata6* deletion specific networks) and PBS *Kras** vs p48Cre (constitutive *Kras* activation specific networks); for situations with two perturbations PBS *Kras**;NT*Gata4*KO vs PBS *Kras** (*Gata4* deletion in a mutated *Kras* context specific networks) and Cae *Kras** vs PBS *Kras** (acute inflammatory insult in a mutated *Kras* context specific networks); and finally, for situations with three perturbations Cae *Kras**;NT*Gata4*KO vs Cae *Kras** (acute inflammatory insult upon *Gata4* deletion in a mutated *Kras* context specific networks).

To isolate the effect of perturbations, an intersection was performed between each pair of networks to obtain the exclusive regulatory relationships between genes of each network that constitute the specific networks. In Venn diagrams, the right regions are those that represent the specific transcriptional network for each pertur-

bation (Figure 2).

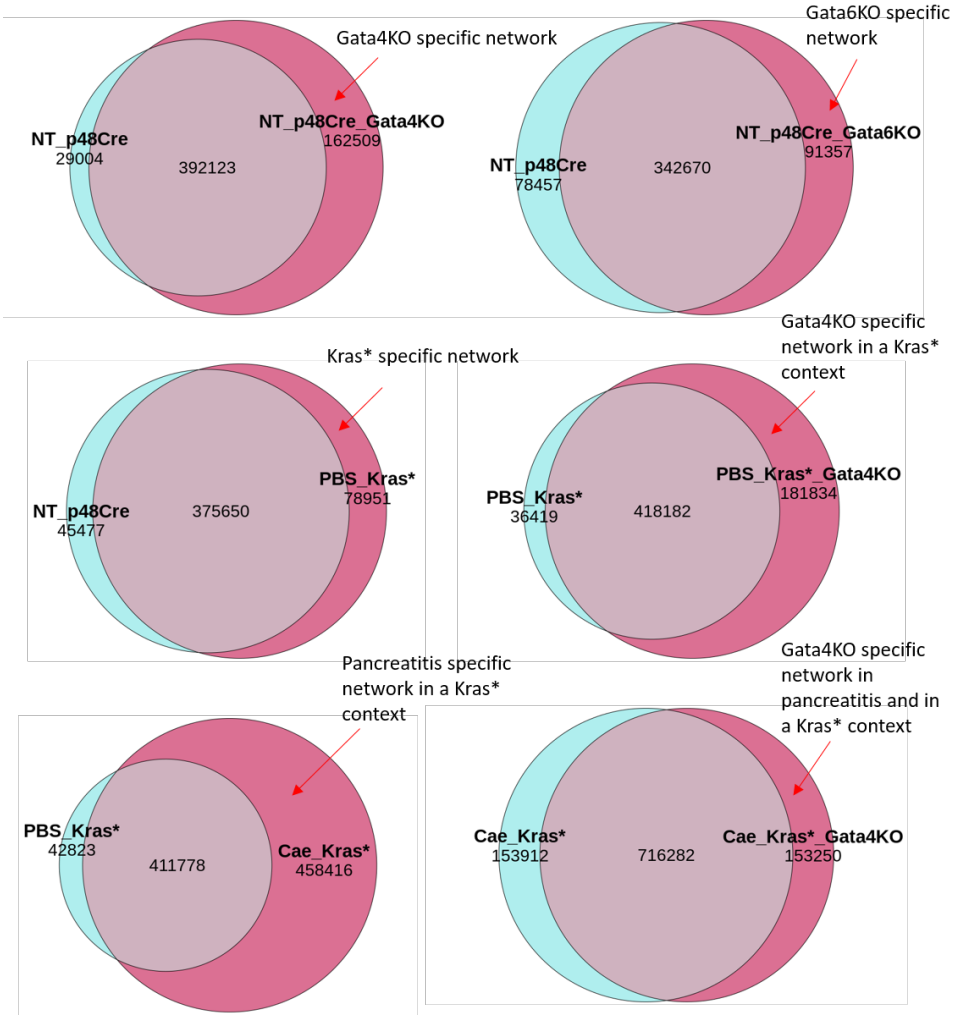


Figure 2: Venn diagrams displaying pairwise comparisons of perturbation specific networks. The region to the right in each diagram represents the specific network for the particular perturbation in each comparison.

2.4 Topological study of subnetworks

The network obtained in the previous steps were very massive and offered a multitude of possibilities to study them. With the aim of extracting data that provide relevant information about the transcriptional regulation involved in acinar differentiation the networks were interrogated by exploring their topology.

Using the *degree*, a topological parameter of the network that defines the number

of connections of each node, the genes that were part of each specific subnet were ranked by the value of this factor. With these values it was studied how the position in the ranking of each gene changed between the control network and the specific network with which it was compared. This process was repeated for each condition.

It is expected that if a gene increases or decreases its position in the list it is because its activity has changed, either positively or negatively, as a consequence of the alteration of the number of OCRs. Therefore, a significant change in positions between the two networks suggests that the transcriptional regulation of acinar identity has changed. For this reason, a threshold of >50 change positions was chosen to take into account those genes whose relevance in the network changed substantially, either because they regulated the expression of a greater number of genes or because of the opposite.

To show the information more clearly, the data was divided into two independent graphs, one for the nodes that had increased positions and the other for the nodes that had decreased positions (Figure 3A).

In the graphs obtained, the list on the left represents the nodes of the network taken as control and the list on the right represents the nodes of the specific network of the perturbation of interest. The nodes appear ranked by their degree within each network. In parentheses, the number of positions that the node has changed in the ranking between both networks is shown. The change of positions and the value of degree was the relevant information in these comparisons. With these data it is possible it is possible to quickly check which nodes are the ones that change their positions the most, that is, those that have gained more relevance in the specific network with perturbation (Figure 3B). The rest of the graphs corresponding to pairwise comparisons can be consulted in Supplementary Figure 3.

2.5 Individual TF subnetworks

After studying the specific network of each perturbation, the analysis of these networks was deepened by extracting subnetworks of certain genes. Subnetworks show all those genes that regulate or are regulated by a specific gene. In the context of this work, the second case was of greater interest, since it allowed to test the impact of one gene on the network.

In order to choose the direct regulatory relationships between genes, which are those that ensure greater confidence that such regulation was occurring, the data set was filtered for those proximal regulatory events. For this purpose, only genes

Figure 3: Venn diagrams for all pairwise comparison where the edges are compared. A) Origin of the data, graph with all position changes, filtering of 50 positions or more and division of the graph according to whether the nodes increase their positions or decrease them. **B)** Close detail of the nodes that increase their position in the pancreatitis specific network in a mutated *Kras* context.

encoding TFs whose TFBSs were located in transcription start site (TSS) regions were chosen for the sake of greater reliability in gene assignment. For the rest of the genes, the TFBSs associated with the TFs they encode were in regions close to the gene but not exactly in the TSS, such as in the intergenic regions or the TTS.

Junb is one of the most highly ranked genes in terms of degree in the pancreatitis-

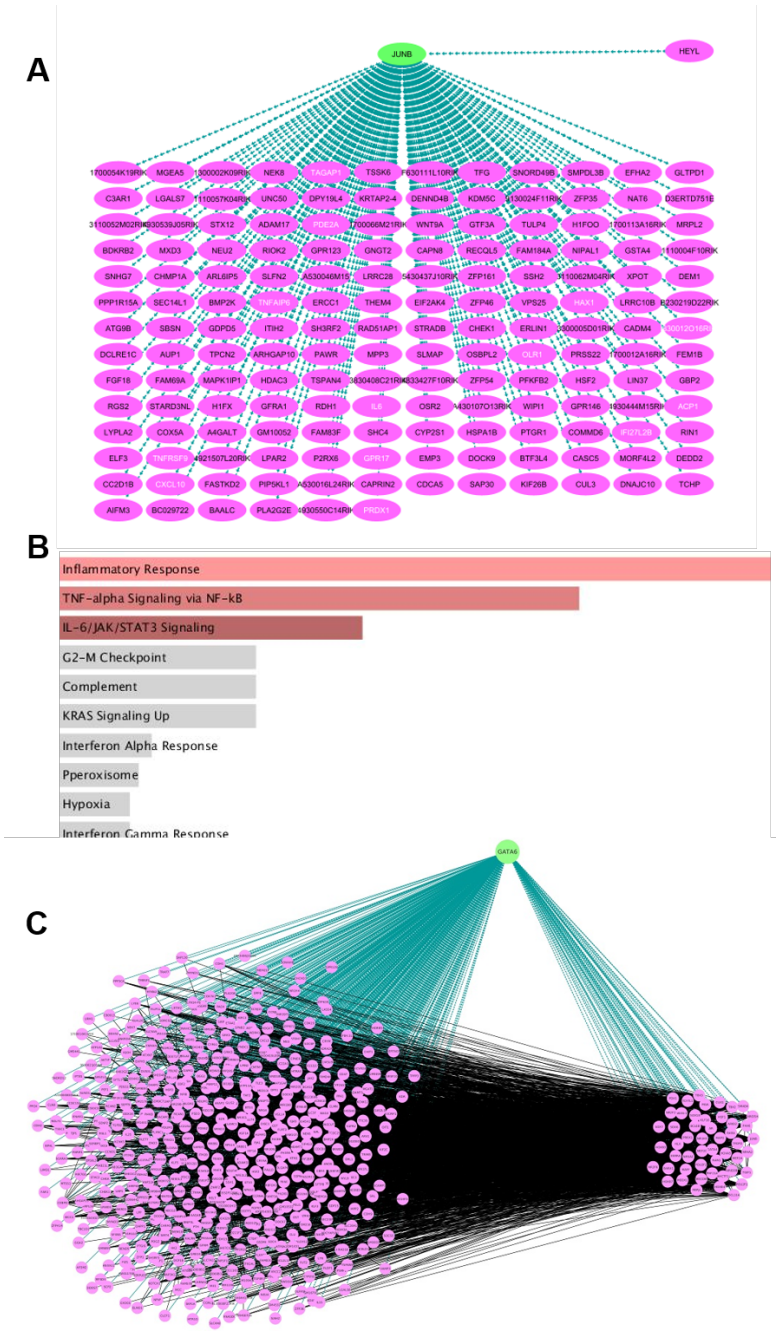


Figure 4: Study of transcriptional networks. A) Network of genes that are regulated or that regulate *Junb* in the specific network of pancreatitis in a mutated *Kras* context. There is only one regulator gene, *Heyl*, the rest are regulated by *Junb*. Unlike the previous networks, the nodes represented in this network were selected because their footprint was found on a TSS region, which ensures a more direct connection. White nodes are genes related to immune system. **B)** Biological functions obtained in Enrichr from genes that are regulated by *Junb*. **C)** Network of genes that are regulated by or that regulate *Gata6* in the specific network of *Gata4*KO. The genes regulated by *Gata6* are on the left and the genes that regulate *Gata6* are on the right.

specific subnetwork (Figure 3B) under the context of mutated *Kras*. It is known to act as a thermostat in the pancreas and its activity is related to pre-inflammatory stages [41]. When analyzing the genes, it is observed that among others it is linked to inflammatory genes which corroborates the function described in the literature (Figure 4A).

To verify that the new genes regulated by *Junb* were indeed involved in the inflammation of the pancreas, a functional analysis was carried out. It was found that the first biological activity described for them was the inflammatory response, as expected (Figure 4B).

This is one example of the type of topological analyses that can be performed on the network and the bidirectional relationship that can be established between the biological results and the information provided by the network. Another possible analysis is the study of the behavior of GATA6 for the specific network of *Gata4*KO since experiments carried out in the research group where this work is developed suggest that a compensatory effect of GATA6 occurs when *Gata4* is knocked out. The network obtained showed that in the specific network of *Gata4*KO, *Gata6* was regulated by more genes and it also regulated more genes (Figure 4C).

2.6 Web application

The two previous analyses are only two examples of how to exploit the data as there are many other alternatives. Given the impossibility of addressing all these options in this work, it was decided to create an interactive tool that would allow to consult the networks in a fast and user-friendly manner. For this purpose, a web application was developed that collects information on the eight specific transcriptional networks previously studied.

The application can be consulted at this link:

`https://jmartinezv.shinyapps.io/Shiny_app/`

It contains the following elements:

- Network selection panel (Figure 5A). By default, the Homeostatic condition is loaded. The remaining options correspond to each perturbation-specific subnetwork. When selecting another dataset, a Venn diagram is shown corresponding to one of the comparisons made between networks. This Venn diagram shows the number of edges for each of the networks (the control network and the network of the perturbation of interest) and for their intersection.

The edges of each perturbation-specific subnetwork are shown in the red portion of the diagram. These are all the options that can be selected with the Network menu:

  - Homeostatic
  - Effect of *Gata4* KO - Context: No treatment
  - Effect of *Gata6* KO - Context: No treatment
  - Effect of mutated *Kras* - Context: PBS
  - Effect of *Gata4* KO - Context: PBS & Mutated *Kras*
  - Effect of caerulein (pancreatitis) - Context: Mutated *Kras*
  - Effect of *Gata4* KO - Context: Caerulein & Mutated *Kras*
  - Effect of caerulein (pancreatitis) - Context: *Gata4* KO & Mutated *Kras*

- TF Degree panel (Figure 5B). Two tables are shown containing the nodes of the control network (left panel) and those of perturbation-specific subnetwork (right panel) ranked by degree. In Homeostatic condition only one panel is shown since no comparison between networks is made.

- Data filtering panel (Figure 5C). These are the parameters with which the table can be filtered:

  - *Footprint*: the footprint from which we extract the motif to which a TF binds. The sequence logo obtained from the PWM is also shown.
  - *Footprint location*: it is the annotated genomic location. A pie chart is shown with the proportion of each of the locations where the TF can bind: Exon, Intergenic, Intron, Promoter-TSS and Transcription Termination Site (TTS).
  - *TFBS score*: This score reflects how well the footprint matches the input TF motif. Once selected, a density plot of the data is displayed.
  - *Footprint score*: This score reflects the quality of the depletion of ATAC-seq signal for a particular footprint. All footprints included passed the default threshold considered by TOBIAS. Once selected a density plot of the data is displayed.
  - *Search box*: to perform any type of search.

- Data table (Figure 5D). Once a perturbation-specific subnetwork is selected displays all the information. The table is dynamically updated when filtered and can be downloaded through the *Download table* button.

- Functional analysis (Figure 6). In order to get biological insights into the queried information, functional enrichment analysis can be done on the target genes using EnrichR package. Output results are displayed as an enrichment plot and table. Different databases can be interrogated:
  - WikiPathways 2019 Mouse
  - KEGG 2019 Mouse
  - MSigDB Hallmark 2020
  - GO Biological Process 2021
  - GO Cellular Component 2021
  - GO Molecular Function 2021
  - ENCODE and ChEA Consensus TFs from ChIP-X
  - ChEA 2016
  - RNAseq Automatic GEO Signatures Mouse Down
  - RNAseq Automatic GEO Signatures Mouse Up

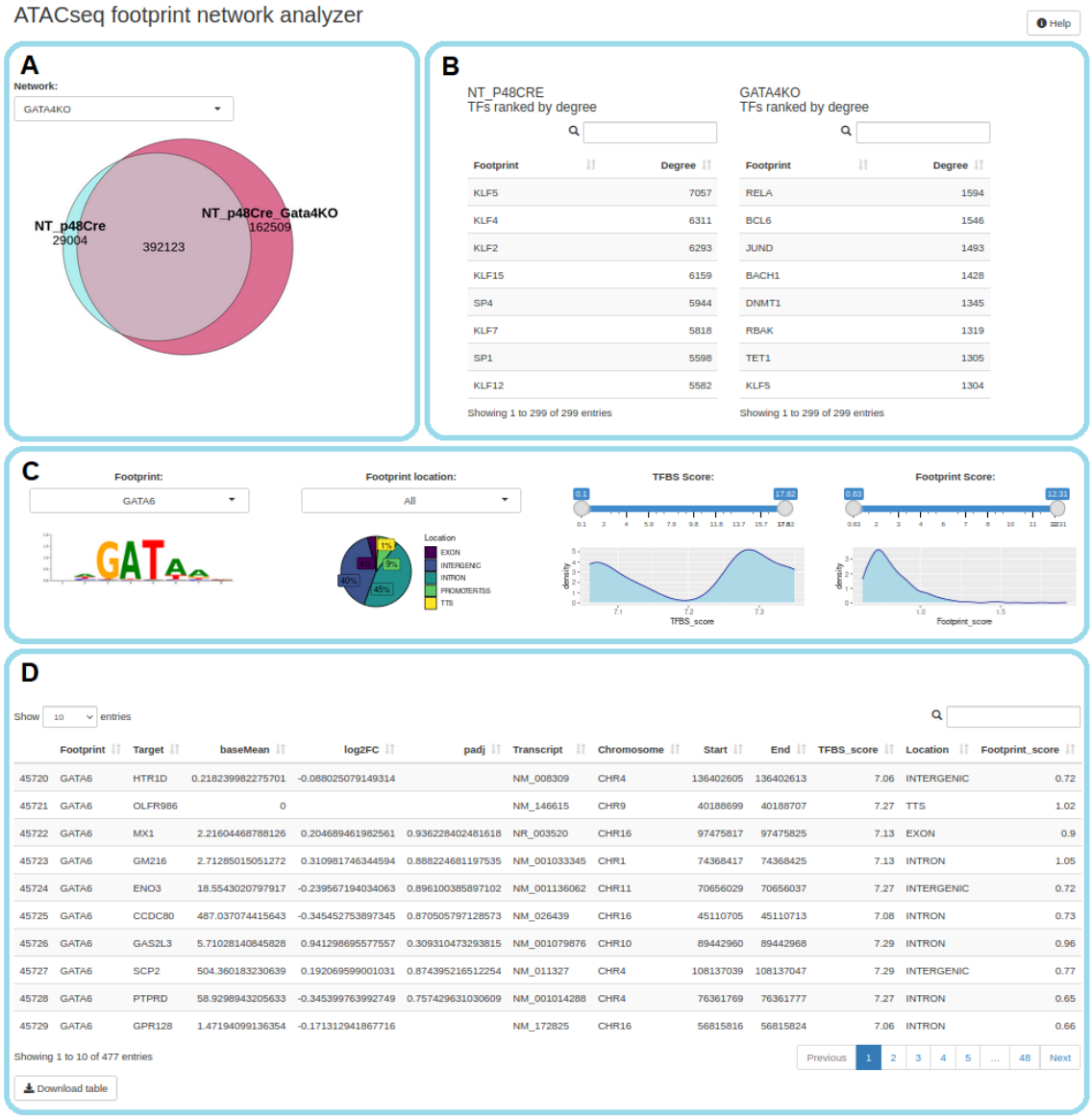


Figure 5: ATAC-seq footprint network analyzer. A) Network selection panel. **B)** TF Degree panel. **C)** Data filtering panel. **D)** Table with all the data.

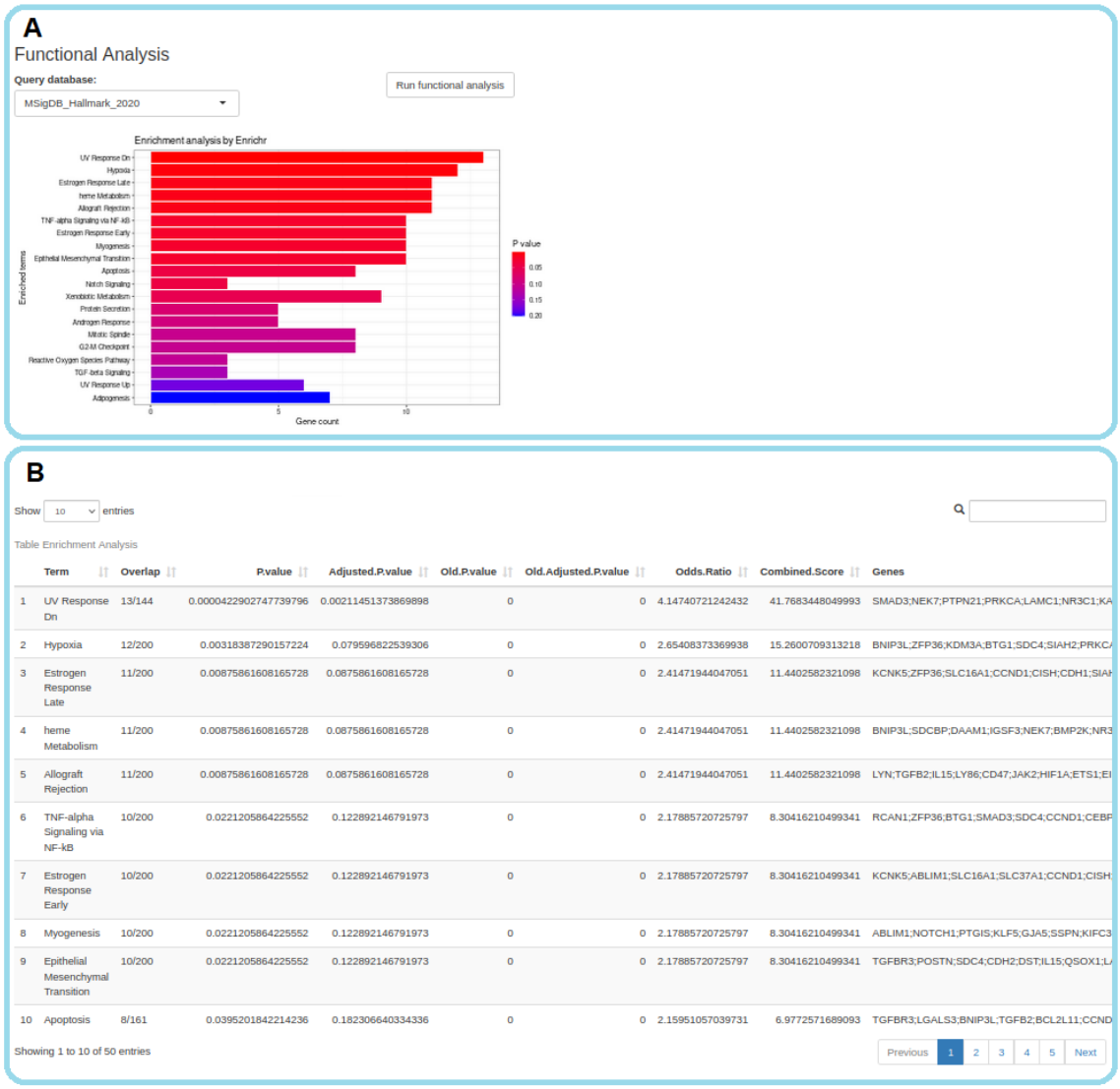


Figure 6: ATAC-seq footprint network analyzer. Functional analysis A) Graphical representation of the functional analysis. **B)** Table with the data of the functional analysis.

3. Discussion

In this work, transcriptional networks of the acinar cells of the mouse pancreas have been obtained both for homeostatic conditions and for scenarios in which the pancreas is challenged, as occurs with the *Gata4*, *Gata6* knockouts, the *Kras* mutation and with the induction of pancreatitis by caerulein treatment. To construct them, ATAC-seq data were used to obtain the OCRs. From these regions, and using the TOBIAS pipeline, the TFBS were obtained and finally the networks were generated. To determine wether the TFs that are specific to pancreatic acinar cells, RNA-seq and scRNA-seq data have been used which, together with the analysis of footprints, made it possible to identify the specific TFBS. With this information, TF-TG networks were generated to model the interactions between regulator genes and the genes that are regulated by them in order to study the regulation of acinar cell identity in mouse pancreas.

The reanalysis of the homeostatic pancreas data allowed to define the process of generation of the transcriptional networks for the perturbation data. The merge of the peaks files of each replicate ensured more robust data without compromising the reliability of the networks. The use of RNA-seq data allowed to identify TFs present in the pancreas while scRNA-seq data served to restrict the TFs to those present only in acinar cells.

There are many ways to interrogate networks and extract information from them. In this work, it has been proposed a topological study of the networks, analyzing the degree of the genes that are part of them and the creation of individual subnetworks for specific regulatory genes such as *Junb* or *Gata6*.

The transcriptional networks of some PDAC precursor perturbations were studied, such as the *Kras* mutation, the *Gata4* and *Gata6* knockout and the presence of pancreatitis. For them, six comparisons were proposed that made it possible to study the effect of each of these perturbations in isolation. In this way, specific networks were obtained for the *Gata4* knockout (both in basal conditions and in the context of mutated *Kras* together with pancreatitis), the *Gata6* knockout (for basal

conditions), the effect of *Kras* mutation and the effect of pancreatitis in the context of mutated *Kras*. The study of the network analysis revealed an upward trend in the number of OCRs to the number of perturbations of each situation. This also implies that a greater number of perturbations would imply networks with more regulatory relationships between genes, which could explain the loss of acinar identity. Among all the perturbations, the presence of pancreatitis is the one that most modifies the topology of the networks, with a significant increase in the OCRs compared to the homeostatic network or those corresponding to the rest of the perturbations. This may reflect not only the changes in the activity of the networks in acinar cells but also changes in the cellular composition of the tissue analysed.

The information that can be extracted from the networks is wide-ranging, since individual genes can be studied observing which genes regulate this individual gene in particular or if it is a gene that regulates others genes. The regulatory relationships of a certain gene can also be studied in the context of a specific network in which a related gene is knockout, as occurs with *Gata6* and *Gata4* respectively. Another possibility is to study the networks to link them with results obtained experimentally, verifying the biological function of the genes that are regulated by a specific gene, as has been done in this work with *Junb*.

Apart from the aforementioned ways of interrogating networks, the manner of analyzing them topologically can also be varied. The study conducted here has considered two different approaches. On the one hand, an unbiased analysis, ranking the genes by degree and checking how their positions change in the ranking between two determined networks. And on the other hand, a biased analysis, interrogating the network for specific genes, also using a proximal regulation on this occasion.

A web application has been developed to offer a simple and clear way of accessing information. Its purpose is to serve as a query tool and to simplify access to data on the transcriptional networks obtained. In addition, functionalities can be further added according to the feedback received. In fact, as a future work, it is expected to add information from RNA-seq data obtained in the same experiment as the ATAC-seq data used to show which genes are up-regulated or down-regulated, offering an even more complete view of transcriptional activity.

However, this study also presents some limitations in the process followed to obtain the transcriptional networks.

- The ATAC-seq signal is less sensitive than ChIP-seq signal because it is unbiased, that is, it does not focus on a protein but on all those that leave a footprint. ChipSeq relies on the quality and specificity of the antibodies and this can vary from one TF to another. This explains why in Figure 3 there are

a large number of regions of the ChIP-seq data that do not coincide with the ATAC-seq data. ATACseq is unbiased in this regard although its analysis does not return direct evidence but inferred information. Despite this limitation, supplementary Figure 1 shows how the ATAC-seq data are enriched not only in the areas that coincide with ChIP-seq for the motif in question, but also in the areas that are outside the overlap. This indicates that the ATAC-seq signal, although it does not allow to detect as many binding regions as ChIP-seq, its results are reliable. On the other hand, the ChIP-seq signal could also be altered by experimental variations which would explain why certain regions defined by ATAC-seq do not intersect with the ChIP-seq data.

- The analysis of the footprints carried out is an inference from the signal depletion observed in the ATAC-seq data and the fit of the motifs with the footprints. This implies that the data obtained is a statistical estimate and must be considered when drawing conclusions. One way to mitigate this limitation is to be more restrictive in terms of TFBS score.

- Due to the lower sensitivity of ATAC-seq, motif sequence similarity might produce assignment of footprints to different TFs of the same family. In this case protein specific resolution might not be possible and results should be TF family considered.

- Different cell populations have been used between the homeostasis data and the rest of the data. The former came from a disaggregation of the whole pancreatic tissue and the latter from FACS sorted epithelial cells. Therefore, this must be taken into account when making comparisons between both experiments.

- The conclusions obtained from the networks need validation by orthologous techniques. In this regard integration with bulk RNA-seq, scRNA-seq and scATAC-seq (data are available for the same perturbation data set studied in this work) is needed.

Despite the limitations, the generated networks are a source of information that can be interrogated in a multitude of different ways. It can therefore serve as a resource both to raise new hypotheses and discover new regulatory mechanisms. In addition, it can also work in the opposite direction, serving as a confirmation tool for results obtained experimentally.

The whole transcriptional network study developed in this master thesis and summarized in the web app tool will serve both to confirm experimental data and, more interestingly, to generate novel hypotheses and therefore draw some light into future lines of research.

4. Methods

4.1 Datasets

For the analysis under homeostatic conditions, a publicly accessible mouse pancreas ATAC-seq from an ATAC-seq atlas [25] was used. 8-week-old male and female C57BL/6J mice were utilized and housed in a pathogen-free, temperature-controlled environment under a 12-hour night/day cycle and were sacrificed by cervical dislocation. The organs were cut into two or three pieces and remained frozen at -80 ºC until the extraction of the nuclei.

The ATAC-seq data from the mouse pancreas that were subjected to some type of perturbation came from a collaboration between Mónica Pérez, from the CNIO, and Scott Lowe's laboratory at Sloan-Kettering Institute, New York. Their contribution was very important for this project so her willingness to help was greatly appreciated. The GEMMs used were *NT p48Cre*: $p48^{+/Cre}$, *NT Gata4KO*: $p48^{+/Cre}$; $Gata4^{lox/lox}$, *NT Gata6KO*: $p48^{+/Cre}$;$Gata6^{lox/lox}$, *PBS/CAE Kras**: $p48^{+/Cre}$; $Kras^{+/LSL\text{-}G12Vgeo}$ and *PBS Kras*; Gata4KO:* $p48^{+/Cre}$; $Kras^{+/LSL\text{-}G12Vgeo}$; $Gata4^{lox/lox}$. The mice used were sacrificed at 10-12 weeks. PBS or cerulein treatment to induce pancreatitis consisted of eight hourly intraperitoneal injections of 80 µg/kg of the CCK analogue caerulein (Bachem) or PBS for two consecutive days. Mice were sacrificed four days after the first injection of caerulein/PBS by CO_2 inhalation.

4.2 ATAC-seq analysis

Paired-end raw FASTQ files were analyzed using the ENCODE ATAC-seq pipeline. To execute the pipeline from FASTQ to peak calling, Caper (Cromwell Assisted Pipeline ExecutoR) was used with the following instruction:

```
caper run [WDL script] -i
```

```
[Input: JSON file containing information of genomic data files]
```

With the Cutadapt v2.5 tool [42] the adapter sequences were eliminated. Next, Bowtie2 v2.3.4.3 [43] was used to map the reads to the reference genome (mm10, GRCm38, December 2011) obtaining the SAM (Sequence Alignment Map) files. These files were transformed into the BAM format with SAMtools v1.9 [44]. The reads that met some of the following characteristics were located and eliminated with the Sambamba v0.6.6 [45] tool: not being mapped, not forming a primary alignment, being duplicated or being mapped to mitochondrial DNA (chrM). PCR duplicates were removed with Picard's MarkDuplicates [46]. Finally, the accessible regions were defined by peak calling using MACS2 [47] and the resulting peaks were those that appeared in all replicates for a threshold of 0.05 IDR.

Quality control showed that the results were reliable. The two PCR bottleneck coefficients, PBC1 and PBC2 [48], were studied. PBC1 shows the ratio between genomic locations where a read is uniquely mapped and locations to which some read maps uniquely. PBC2 is the ratio between the number of genomic locations where only one read maps uniquely and the number of genomic locations where two reads map uniquely. The replicates also passed the TSS enrichment threshold in OCRs for the mm10 genome.

4.3 Footprinting analysis

The footprinting analysis was performed with the TOBIAS toolkit. The ATACorrect module was utilized to correct the readings taking into account the bias introduced by Tn5. The ScoreBigwig function was used to calculate the footprint scores of cutsites across accessible regions. With BINDetect, the TF binding events were studied from the footprints and information on the motifs. A threshold p-value of 0.001 was chosen to differentiate between TF bound or unbound. With the CreateNetwork function, the associations between the TF bounds and the target genes were modeled. Using the results returned in this last step, the transcriptional networks were created.

4.4 Peak merging and annotation

The peaks files of the homeostatic pancreas analysis were merged, those of male on the one hand and those of female on the other, to obtain a consensus file of peaks to obtain more robust OCRs. The peaks files of the perturbed pancreas analysis

were also merged and in this case a merge was performed for each perturbation data set. This process was carried out using the mergePeaks function of the HOMER software [28]. The *-d given* option was used to ensure literal overlaps between the peaks of each replicate.

Peak annotation was done with the HOMER annotatePeaks.pl function. The reference genome used was mm10 (GRCm38, December 2011). Peaks were annotated from a gtf annotation file obtained from the UCSC Genome Browser [49].

With the HOMER findMotifsGenome.pl function, motif analysis was performed to locate enriched motifs both in the ChIP-seq signal and in the data obtained from ATAC-seq.

4.5 Transcription factor binding motifs

Information regarding TF binding motifs was obtained from CIS-BP. The motifs of TEAD2, SOX9, RBPJL, NFYB, FOXA3 and ETS2 were obtained from JASPAR CORE 2020 as they were not present in CIS-BP.

The motifs were restricted to those corresponding to the TFs expressed in acinar cells. RNA-seq data were used to select them and TFs with an expression level greater than 3 RPKM were chosen. ScRNA-seq data was utilized to limit motif information to those TFs with $\geq$1 acinar cell expression.

The motif files were manipulated in R with Bioconductor Universal motif package [50].

4.6 Network visualization and analysis

Venn diagrams, density charts and bump charts were calculated with RStudio v1.1.419 [51].

The networks were represented with Cytoscape v3.8.2 [52]. The calculation of the intersections and differences between networks were also made with this tool using the *merge* functionality of its options menu. Within this function, to obtain the portion of the network of interest when performing the difference between networks, it is necessary to place the network of interest first and the control network second. To obtain subnets of a larger network from a desired node (or nodes) it was necessary to follow the following menu path: *Select desired nodes >File >New network >From*

selected nodes-All edges. The previous process returned all the connections with the node of interest, both in those cases in which it was an incoming node (regulated gene) and an outgoing node (regulatory gene). To keep only the cases in which the node of interest is outgoing, the following was done: *The node of interest is selected in the subnetwork >Select >Nodes >First Neighbors of Selected Nodes >Directed: outgoing*. The case for incoming nodes is analogous. The ranking by degree of the network nodes was carried out with the cytoHubba [53] plugin integrated in Cytoscape. The biological functions from a set of genes were obtained with the Enrichr online tool [54].

4.7 Web application development

For the development of the application, the R Shiny package [55] was used. This package contains the necessary tools and functions to create a dashboard in which to display the information that the developer wants.

The code is split between *ui* and *server* blocks. The first defines the aspects related to the user interface and the graphical aspects of the application. The second describes the logic behind the functions performed by the application.

Filters like slider inputs or select inputs are provided by the Shiny package. The table is defined with the DataTable class from the DT package. The application makes use of reactive programming to update the displayed information in real time. For the functional analysis, the enrichR package was used, which contains functions that allow access to the content of the Enrichr online tool. For the pie chart and the density plots, the *ggplot2* package was used. The web has been organized using a 12-column grid system. For specific modifications of the visual design of the web not supported by Shiny, direct calls to HTML were used.

The website was published on the internet using the Shinyapp.io hosting service provided by Rstudio.

4.8 Code availability

The code executed to carry out the footprinting analysis and the generation of the graphs can be consulted at the following link:

`https://github.com/FranSoriano/BioinfoTFM`

Bibliography

[1] Hezel, A. F., Kimmelman, A. C., Stanger, B. Z., Bardeesy, N., & Depinho, R. A. (2006). Genetics and biology of pancreatic ductal adenocarcinoma. *Genes & development, 20*(10), 1218–1249.

[2] Pour, P. M., Pandey, K. K., & Batra, S. K. (2003). What is the origin of pancreatic adenocarcinoma? *Molecular cancer, 2*, 13.

[3] Sarantis, P., Koustas, E., Papadimitropoulou, A., Papavassiliou, A. G., & Karamouzis, M. V. (2020). Pancreatic ductal adenocarcinoma: Treatment hurdles, tumor microenvironment and immunotherapy. *World journal of gastrointestinal oncology, 12*(2), 173–181.

[4] Bengtsson, A., Andersson, R., & Ansari, D. (2020). The actual 5-year survivors of pancreatic ductal adenocarcinoma based on real-world data. *Scientific reports, 10*(1), 16425.

[5] Aguirre, A. J., Bardeesy, N., Sinha, M., Lopez, L., Tuveson, D. A., Horner, J., Redston, M. S., & DePinho, R. A. (2003). Activated Kras and Ink4a/Arf deficiency cooperate to produce metastatic pancreatic ductal adenocarcinoma. *Genes & development, 17*(24), 3112–3126.

[6] Hingorani, S. R., Petricoin, E. F., Maitra, A., Rajapakse, V., King, C., Jacobetz, M. A., Ross, S., Conrads, T. P., Veenstra, T. D., Hitt, B. A., Kawaguchi, Y., Johann, D., Liotta, L. A., Crawford, H. C., Putt, M. E., Jacks, T., Wright, C. V., Hruban, R. H., Lowy, A. M., & Tuveson, D. A. (2003). Preinvasive and invasive ductal pancreatic cancer and its early detection in the mouse. *Cancer cell, 4*(6), 437–450.

[7] Habbe, N., Shi, G., Meguid, R. A., Fendrich, V., Esni, F., Chen, H., Feldmann, G., Stoffers, D. A., Konieczny, S. F., Leach, S. D., & Maitra, A. (2008). Spontaneous induction of murine pancreatic intraepithelial neoplasia (mPanIN) by acinar cell targeting of oncogenic Kras in adult mice. *Proceedings of the National Academy of Sciences of the United States of America, 105*(48), 18913–18918.

[8] Gidekel Friedlander, S. Y., Chu, G. C., Snyder, E. L., Girnius, N., Dibelius, G., Crowley, D., Vasile, E., DePinho, R. A., & Jacks, T. (2009). Context-dependent transformation of adult pancreatic cells by oncogenic K-Ras. *Cancer cell, 16*(5), 379–389.

[9] Wang, L., Xie, D., & Wei, D. (2019). Pancreatic Acinar-to-Ductal Metaplasia and Pancreatic Cancer. *Methods in molecular biology (Clifton, N.J.), 1882*, 299–308.

[10] Stanger, B. Z., & Hebrok, M. (2013). Control of cell identity in pancreas development and regeneration. *Gastroenterology, 144*(6), 1170–1179.

[11] Martinelli, P., Madriles, F., Cañamero, M., Pau, E. C., Pozo, N. D., Guerra, C., & Real, F. X. (2016). The acinar regulator Gata6 suppresses KrasG12V-driven pancreatic tumorigenesis in mice. *Gut, 65*(3), 476–486.

[12] Dassaye, R., Naidoo, S., & Cerf, M. E. (2016). Transcription factor regulation of pancreatic organogenesis, differentiation and maturation. *Islets, 8*(1), 13–34.

[13] Slack J. M. (2007). Metaplasia and transdifferentiation: from pure biology to the clinic. *Nature reviews. Molecular cell biology, 8*(5), 369–378.

[14] Wong, C. H., Li, Y. J., & Chen, Y. C. (2016). Therapeutic potential of targeting acinar cell reprogramming in pancreatic cancer. *World journal of gastroenterology, 22*(31), 7046–7057.

[15] Margolin, A. A., Wang, K., Lim, W. K., Kustagi, M., Nemenman, I., & Califano, A. (2006). Reverse engineering cellular networks. *Nature protocols, 1*(2), 662–671.

[16] Davidson, E. H., & Erwin, D. H. (2006). Gene regulatory networks and the evolution of animal body plans. *Science (New York, N.Y.), 311*(5762), 796–800.

[17] Redhu, N., & Thakur, Z. (2022). Network biology and applications. In *Bioinformatics* (pp. 381-407). Academic Press.

[18] Zhang, B., Tian, Y., & Zhang, Z. (2014). Network biology in medicine and beyond. *Circulation. Cardiovascular genetics*, 7(4), 536–547.

[19] Madriles, F. (2017). Role of gata4 in pancreatic physiology and carcinogenesis. Ph.D. Thesis, Universidad Autónoma de Madrid.

[20] Gong, Y., Zhang, L., Zhang, A., Chen, X., Gao, P., & Zeng, Q. (2018). GATA4 inhibits cell differentiation and proliferation in pancreatic cancer. *PloS one, 13*(8), e0202449.

[21] Martinelli, P., Carrillo-de Santa Pau, E., Cox, T., Sainz, B., Jr, Dusetti, N., Greenhalf, W., Rinaldi, L., Costello, E., Ghaneh, P., Malats, N., Büchler, M., Pajic, M., Biankin, A. V., Iovanna, J., Neoptolemos, J., & Real, F. X. (2017). GATA6 regulates EMT and tumour dissemination, and is a marker of response to adjuvant chemotherapy in pancreatic cancer. *Gut, 66*(9), 1665–1676.

[22] Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods, 10*(12), 1213–1218.

[23] Buenrostro, J. D., Wu, B., Chang, H. Y., & Greenleaf, W. J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current protocols in molecular biology, 109*, 21.29.1–21.29.9.

[24] Galas, D. J., & Schmitz, A. (1978). DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic acids research, 5*(9), 3157–3170.

[25] Liu, C., Wang, M., Wei, X., Wu, L., Xu, J., Dai, X., Xia, J., Cheng, M., Yuan, Y., Zhang, P., Li, J., Feng, T., Chen, A., Zhang, W., Chen, F., Shang, Z., Zhang, X., Peters, B. A., & Liu, L. (2019). An ATAC-seq atlas of chromatin accessibility in mouse tissues. *Scientific data, 6*(1), 65.

[26] Anshul, K., Nathan, B., Daniel, K., Chuan Sheng, F. & Lee, J. (2016) ENCODE ATAC-seq pipeline. ENCODE-DCC, GitHub repository https://github.com/ENCODE-DCC/atac-seq-pipeline.

[27] Li, Q., Brown, J. B., Huang, H., & Bickel, P. J. (2011). Measuring reproducibility of high-throughput experiments. *The annals of applied statistics, 5*(3), 1752-1779.

[28] Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell, 38*(4), 576–589.

[29] Bentsen, M., Goymann, P., Schultheis, H., Klee, K., Petrova, A., Wiegandt, R., Fust, A., Preussner, J., Kuenne, C., Braun, T., Kim, J., & Looso, M. (2020). ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nature communications, 11*(1), 4267.

[30] Koohy, H., Down, T. A., & Hubbard, T. J. (2013). Chromatin accessibility data sets show bias due to sequence specificity of the DNase I enzyme. *PloS one, 8*(7), e69853.

[31] Karabacak Calviello, A., Hirsekorn, A., Wurmus, R., Yusuf, D., & Ohler, U. (2019). Reproducible inference of transcription factor footprints in ATAC-seq and DNase-seq datasets using protocol-specific bias modeling. *Genome biology, 20*(1), 42.

[32] He, H. H., Meyer, C. A., Hu, S. S., Chen, M. W., Zang, C., Liu, Y., Rao, P. K., Fei, T., Xu, H., Long, H., Liu, X. S., & Brown, M. (2014). Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nature methods, 11*(1), 73–78.

[33] Li, Z., Schulz, M. H., Look, T., Begemann, M., Zenke, M., & Costa, I. G. (2019). Identification of transcription factor binding sites using ATAC-seq. *Genome biology, 20*(1), 45.

[34] Siddharthan R. (2010). Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. *PloS one, 5*(3), e9722.

[35] Melendez, E., Chondronasiou, D., Mosteiro, L., Martínez de Villarreal, J., Fernández-Alfara, M., Lynch, C. J., Grimm, D., Real, F. X., Alcamí, J., Climent, N., Pietrocola, F., & Serrano, M. (2022). Natural killer cells act as an extrinsic barrier for in vivo reprogramming. *Development (Cambridge, England), 149*(8), dev200361.

[36] Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J. C., Galli, M., Lewsey, M. G., Huang, E., Mukherjee, T., Chen, X., Reece-Hoyes, J. S., ... Hughes, T. R. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell, 158*(6), 1431–1443.

[37] Fornes, O., Castro-Mondragon, J. A., Khan, A., van der Lee, R., Zhang, X., Richmond, P. A., Modi, B. P., Correard, S., Gheorghe, M., Baranašić, D., Santana-Garcia, W., Tan, G., Chèneby, J., Ballester, B., Parcy, F., Sandelin, A., Lenhard, B., Wasserman, W. W., & Mathelier, A. (2020). JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic acids research, 48*(D1), D87–D92.

[38] Herzenberg, L. A., Parks, D., Sahaf, B., Perez, O., Roederer, M., & Herzenberg, L. A. (2002). The history and future of the fluorescence activated cell sorter and flow cytometry: a view from Stanford. *Clinical chemistry, 48*(10), 1819–1827.

[39] Westphalen, C. B., & Olive, K. P. (2012). Genetically engineered mouse models of pancreatic cancer. *Cancer journal (Sudbury, Mass.), 18*(6), 502–510.

[40] Fan, Z., Fan, K., Yang, C., Huang, Q., Gong, Y., Cheng, H., Jin, K., Liu, C., Ni, Q., Yu, X. & Luo, G. (2018). Critical role of KRAS mutation in pancreatic ductal adenocarcinoma. *Translational Cancer Research, 7*(6), 1728-1736.

[41] Cobo, I., Martinelli, P., Flández, M., Bakiri, L., Zhang, M., Carrillo-de-Santa-Pau, E., Jia, J., Sánchez-Arévalo Lobo, V. J., Megías, D., Felipe, I., Del Pozo, N., Millán, I., Thommesen, L., Bruland, T., Olson, S. H., Smith, J., Schoonjans, K., Bamlet, W. R., Petersen, G. M., Malats, N., ... Real, F. X. (2018). Transcriptional regulation by NR5A2 links differentiation and inflammation in the pancreas. *Nature, 554*(7693), 533–537.

[42] Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal, 17*(1), 10-12.

[43] Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods, 9*(4), 357–359.

[44] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England), 25*(16), 2078–2079.

[45] Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., & Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. *Bioinformatics (Oxford, England), 31*(12), 2032–2034.

[46] Broad Institute. Picard toolkit: A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF. (2019) Broad Institute, GitHub repository https://github.com/broadinstitute/picard.

[47] Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., & Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome biology, 9*(9), R137.

[48] ENCODE project web (2021). Terms and definitions. https://www.encodeproject.org/data-standards/terms/.

[49] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. *Genome Res.* 2002 Jun;12(6):996-1006.

[50] Tremblay BJ (2021). universalmotif: Import, Modify, and Export Motifs with R. R package version 1.12.1, https://bioconductor.org/packages/universalmotif/.

[51] RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. https://www.rstudio.com/.

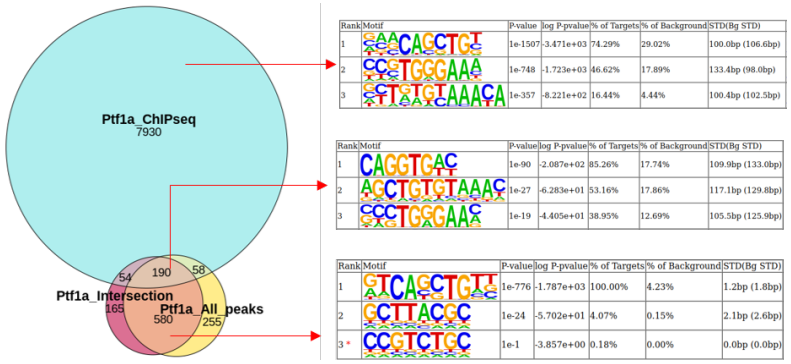
[52] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research, 13*(11), 2498–2504.

[53] Chin, C. H. et al. cytoHubba: Identifying hub objects and sub-networks from complex interactome. *BMC Syst. Biol. 8*, S11 (2014).

[54] Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., & Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics, 14*, 128.

[55] Chang, W., Cheng, Joe., Allaire, J. J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert,A., & Borges, B. (2021). shiny: Web Application Framework for R. R package version 1.7.1. https://CRAN.R-project.org/package=shiny.

Supplementary Information

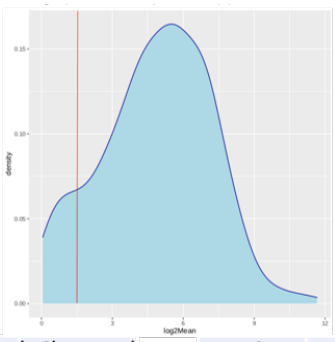


Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)
1		1e-1507	-3.471e+03	74.29%	29.02%	100.0bp (106.6bp)
2		1e-748	-1.723e+03	46.62%	17.89%	133.4bp (98.0bp)
3		1e-357	-8.221e+02	16.44%	4.44%	100.4bp (102.5bp)

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)
1		1e-90	-2.087e+02	85.26%	17.74%	109.9bp (133.0bp)
2		1e-27	-6.283e+01	53.16%	17.86%	117.1bp (129.8bp)
3		1e-19	-4.405e+01	38.95%	12.69%	105.5bp (125.9bp)

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)
1		1e-776	-1.787e+03	100.00%	4.23%	1.2bp (1.8bp)
2		1e-24	-5.702e+01	4.07%	0.15%	2.1bp (2.6bp)
3 *		1e-1	-3.857e+00	0.18%	0.00%	0.0bp (0.0bp)

Figure 1: Motif analysis of Ptf1a. Although the overlap in this case is not very good, the motif analysis shows how the ATAC-seq signal is enriched in Ptf1a in all its regions.



Transc. factor	Mean expression	log2(mean exp.)
Xbp1	3224,526308	11,65487152
Bhlha15	2117,083587	11,04786252
Rbpjl	1547,516119	10,59573872
Atf4	974,2099003	9,928088834
Atf5	926,9699548	9,856378768
Tead2	817,1204197	9,674404895
Atf6	553,6087662	9,112722977
Klf15	504,1595133	8,977736457
Jund	493,933983	8,94817442
Cenpb	423,8887302	8,7275418
Nfe2l1	399,7284174	8,64287633
Myrf	359,1894136	8,48860102
Cxxc1	354,1297566	8,468134265
Cux1	319,236039	8,318479718
Tef	290,7172586	8,18347291
Nr3c1	288,3094276	8,171474203
Klf9	287,1549494	8,165685618
Gata4	285,0187013	8,154912773
Irf6	284,257066	8,151052401
Ptf1a	280,5579125	8,132154791
Rxra	276,7830787	8,112611935
Nfic	272,7066766	8,091206212
Stat3	272,4564463	8,089881815
Creb3l1	271,5448377	8,085046626
Tfdp2	270,3198562	8,078523678
Srebf1	261,1951982	8,028984564
Ets2	250,7319208	7,970001868
Dbp	242,5844532	7,922343284
Usf2	229,6053672	7,843012556
Nr2f6	215,9091004	7,754280243
Mlxip	215,2683238	7,749992236
Nr5a2	209,2595468	7,709149633
Hbp1	209,1492885	7,70838928
Ubp1	191,192045	7,578878688
Kdm2a	181,1112013	7,500731966
Nfat5	178,6499606	7,480991785
Cebpg	177,323804	7,470242407
Cebpa	176,967411	7,467339899
Creb3	172,3834591	7,429477538
Stat6	168,1824841	7,393883649
Sp1	161,2911402	7,333523382
Meis2	160,2990908	7,324622432
Rela	153,9129852	7,265971142
Mef2d	147,6892815	7,206421316
Hnf4a	146,6085129	7,195825067
Mlx	143,1340665	7,16122327
Foxp4	142,5768717	7,155596162
Ahctf1	142,4538485	7,154350788
Rxrb	141,9307585	7,149043466
Cic	140,7384571	7,136872792

Transc. factor	Mean expression	log2(mean exp.)
Myc	140,1365481	7,130689454
Bach1	137,8030207	7,106463703
Irf3	136,9386647	7,097386039
Hif1a	134,9045479	7,075795175
Smad5	134,6394427	7,0729573
Clock	132,8726914	7,053900816
Cdc5l	132,0732135	7,045194085
Tcf4	131,8396451	7,042640454
Ctcf	128,2050977	7,002309817
Crebzf	127,5420719	6,994829412
Srebf2	126,4076908	6,981940431
Usf1	124,2997116	6,957679138
Tet3	123,9792793	6,953955212
Smad3	121,8193594	6,928599613
Smad4	121,629616	6,926350748
Arnt	120,2248637	6,909591481
Atf6b	120,0736556	6,907775846
Foxa3	118,616909	6,890165872
Foxo4	116,8873599	6,868975117
Atf1	116,1859958	6,860292378
Tcf12	115,5889886	6,852860158
Nfix	114,8319386	6,843380149
Jun	114,5779435	6,840185539
Yy1	113,4415131	6,825804871
Spdef	112,6590505	6,815819407
Rbpj	112,4869298	6,813613569
Foxa2	110,1344251	6,783121678
Mga	108,9999808	6,768184071
Nfe2l2	107,9715197	6,754507004
Etv6	107,9553845	6,754291392
Max	107,6877639	6,750710521
Rreb1	107,330078	6,74591062
Foxo1	105,5633012	6,721964563
Pbx1	104,4675258	6,706910734
Ehf	104,0446303	6,7010587
Gabpa	103,607476	6,694984298
Mafk	95,30280702	6,574446802
Mecp2	91,36471321	6,513565171
Srf	89,38368428	6,481939607
Elf2	88,58452049	6,468982715
Sox9	85,87053797	6,424091325
Creb3l2	85,86176496	6,423943924
Atf2	85,84711341	6,42369772
Arid2	85,55762243	6,418824487
Bhlhe40	84,27105201	6,39696523
Gata6	79,74670101	6,317352933
Foxj3	78,71970787	6,298652962
Plagl1	75,49378317	6,23828594
Creb1	75,21732334	6,232993063
Hmg20b	71,25590861	6,154937744

Figure 2: Density plot of gene expression in acinar cells and table with first top 100 expressed TFs.

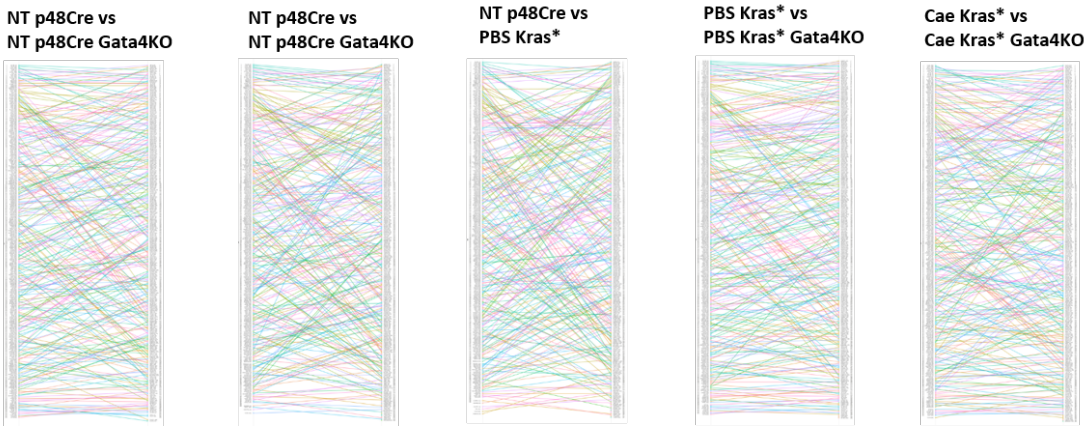


Figure 3: Variation of the position occupied by the TFs in the lists ranked by their degree for the pairwise comparison between networks. Graphs analogous to those in Figure 3A for the rest of the comparisons between networks carried out in this work.

XBP1	SREBF1	MYC	RBPJ	TFE3	TERF2	NFKB2	KLF16	THAP1	HES7	ERG	NPAS2
BHLHA15	ETS2	BACH1	FOXA2	ESRRA	FOXJ2	TRP53	DDIT3	HSF2	RORA	ETV1	SNAI1
RBPJL	DBP	IRF3	MGA	STAT1	HEYL	POU2F1	SP2	NKX2-2	ELK1	E2F3	MESP2
ATF4	USF2	HIF1A	NFE2L2	HES6	NR2F2	KDM2B	LIN54	TCF7	WT1	GLIS3	SOX5
ATF5	NR2F6	SMAD5	ETV6	E2F4	MEIS1	TGIF1	CEBPD	IRF7	HOMEZ	TGIF2	HOXA5
TEAD2	MLXIP	CLOCK	MAX	CREBL2	ETS1	NFYA	MLXIPL	ELF4	FOXO6	IRF4	AR
ATF6	NR5A2	CDC5L	RREB1	FOXK1	SOX6	GLIS2	RFX5	ELK4	ATF7	NAIF1	NKX6-1
KLF15	HBP1	TCF4	FOXO1	RFX7	TCF3	NFIL3	SP100	GMEB1	SP4	EBF3	BCL6B
JUND	UBP1	CTCF	PBX1	FOXP2	ELK3	MNT	OVOL2	TCF7L1	FLI1	ARID3A	HAND2
CENPB	KDM2A	CREBZF	EHF	NFYB	TIGD2	DNMT1	SOX18	E2F2	MYPOP	ATF3	STAT4
NFE2L1	NFAT5	SREBF2	GABPA	GMEB2	SIX5	RFX1	RBAK	TFAP4	JDP2	GLI3	INSM1
MYRF	CEBPG	USF1	MAFK	HNF1B	SOX13	RARG	PRRX1	BCL6	SIX4	NFATC4	KLF7
CXXC1	CEBPA	TET3	MECP2	HSF1	SOX12	JUNB	FOXN3	RELB	TRPS1	KLF5	TCF21
CUX1	CREB3	SMAD3	SRF	IRF2	ETV3	MEF2A	MEF2C	CREM	NFE2L3	SNAI3	TBX3
TEF	STAT6	SMAD4	ELF2	PHF21A	IRF9	MYNN	SETBP1	PPARG	GRHL1	HEY1	POU2F2
NR3C1	SP1	ARNT	SOX9	GFI1	KLF4	MEIS3	TCF7L2	HLF	BHLHE41	PAX6	RARB
KLF9	MEIS2	ATF6B	CREB3L2	HNF1A	SOX4	TFEB	ARID5B	MITF	SP110	TBX2	HOXB4
GATA4	RELA	FOXA3	ATF2	MECOM	EGR1	NR4A2	OVOL1	OSR1	KLF12	KLF8	NEUROD1
IRF6	MEF2D	FOXO4	ARID2	BBX	GRHL2	HES1	MAFB	HIC1	ONECUT1	VDR	MYBL2
PTF1A	HNF4A	ATF1	BHLHE40	MTF1	KLF2	PKNOX1	MAFG	ARID3B	PROX1	TET1	GATA5
RXRA	MLX	TCF12	GATA6	ETV5	FOSL2	PLAGL2	PDX1	SOX7	BARX1	NKX2-3	SNAI2
NFIC	FOXP4	NFIX	FOXJ3	NR1H3	ARNTL	TFCP2	ARID5A	BCL11A	HLX	SOX17	MYB
STAT3	AHCTF1	JUN	PLAGL1	ELF3	PRDM4	KLF11	TBP	IRF5	RFX2	PRDM1	HOXB3
CREB3L1	RXRB	YY1	CREB1	FOXN2	CEBPB	RARA	RFX3	REST	FOXJ1	RUNX1	HOXB7
TFDP2	CIC	SPDEF	HMG20B	STAT2	NR2C1	HMBOX1	CUX2	MAFF	MEOX1	FOXM1	

Table 1: The 299 TFs selected, from TFs expressed in acinar cells, as input for the building of transcription factor networks.