UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

TRABAJO FIN DE MÁSTER

Deciphering a gene regulation network in normal mouse pancreas through a multi-omic integrative approach

Máster Universitario en Bioinformática y Biología Computacional

Autor: Pérez Martínez, Pablo

Tutores: Martínez de Villareal, Jaime
Real Arribas, Francisco X

Ponente: Díaz Uriarte, Ramón

Centro Nacional de Investigaciones Oncológicas (CNIO)

Departamento de Bioquímica UAM

Junio de 2021

Acknowledgments

First of all, I would like to thank Francisco X. Real for the great opportunity to work in such a competent group and for his guidance and always helpful teachings on the complex biology of the pancreas.

I would like to especially thank Jaime Martínez de Villareal for his unvaluable support and guidance in the development of this project under unusual COVID-19 conditions and for his closeness, making me feel part of the group from the first day.

I am also grateful to Arnau Sebe Pedrós, whose suggestions have been key to determine the course of this project.

Finally, I would like to thank the entire Epithelial Carcinogenesis group at CNIO and, especially, Mark and Mónica for their interest and enriching discussions and ideas regarding this work.

Index

Abbreviations

- **ADM**: Acinar-to-Ductal-Metaplasia
- **ATAC-seq:** Assay for Transposase-Accessible Chromatin using sequencing
- **BAM**: Binary Alignment Map
- **ChIP-seq:** Chromatin InmunoPrecipitation sequencing
- **DWM**: Dinucleotide Weight Matrix
- **GEMM**: Genetically Engineered Mouse Model
- **GRN**: Gene Regulatory Network
- **GTEx**: Genotype-Tissue Expression
- **IDR**: Irreproducible Discovery Rate
- **NGS**: Next-Generation-Sequencing
- **OCR**: Open Chromatin Region
- **PDAC**: Pancreatic Ductal Adenocarcinoma
- **PFM**: Position Frequency Matrix
- **PPM**: Position Probability Matrix
- **RPKM**: Reads Per Kilobase per Million mapped reads
- **scRNA-seq**: single cell RNA sequencing
- **TF**: Transcription Factor
- **TFBS**: Transcription Factor Binding Site
- **TSS**: Transcription Start Site
- **WT**: Wild Type
- **ZFP**: Zinc-Finger Protein

Abstract

Pancreatic acinar cells compose around 85% of the exocrine component of the pancreas, which constitutes the vast majority of the tissue. Genetically Engineered Mouse Models (GEMMs) provide evidence that pancreatic ductal adenocarcinoma (PDAC) can efficiently arise from acinar cells through a transdifferentiation process called acinar-to-ductal-metaplasia (ADM), proposing the loss of acinar cell identity as the predominant origin for PDAC. Here, we present a comprehensive multi-omic integrative approach to generate a network-based resource to interrogate the transcriptional regulation underlying acinar cell identity in wild type (WT) mouse pancreas. As a proof-of-concept, we examine the regulatory activity of several acinar-expressed transcription factors (TFs) involved in pancreas regulation and validate it by comparison with experimental ChIP-seq analysis, obtaining consistent results. We consider that this approach represents a valuable resource to perform *a priori* analyses that can be experimentally validated providing new knowledge to the field. Moreover, the presented methodology will be further explored to determine the optimal parameters for improving the potential in the detection of different regulatory events, and will be applied to GEMMs displaying different conditions, as well as to other organisms like human to cross-validate the results and the usefulness of our resource.

Introduction

The pancreas is an endoderm-derived organ that plays a crucial role in the metabolism of all vertebrates and can be functionally divided into an endocrine and an exocrine component. The endocrine pancreas is composed of five specialized cell types grouped in the Langerhans islets and secrete peptide hormones responsible for glucose homeostasis. The exocrine component represents more than 95% of the pancreatic mass in mammalians and constitutes a branching network of acinar and ductal cells. The acinar cells compose approximately 85% of the exocrine pancreas and produce the hydrolytic digestive enzymes that are secreted into the lumen of ductal cells, which convey the enzymes to the gut for protein, carbohydrate and fat digestion[1,2,3] (Figure 1).

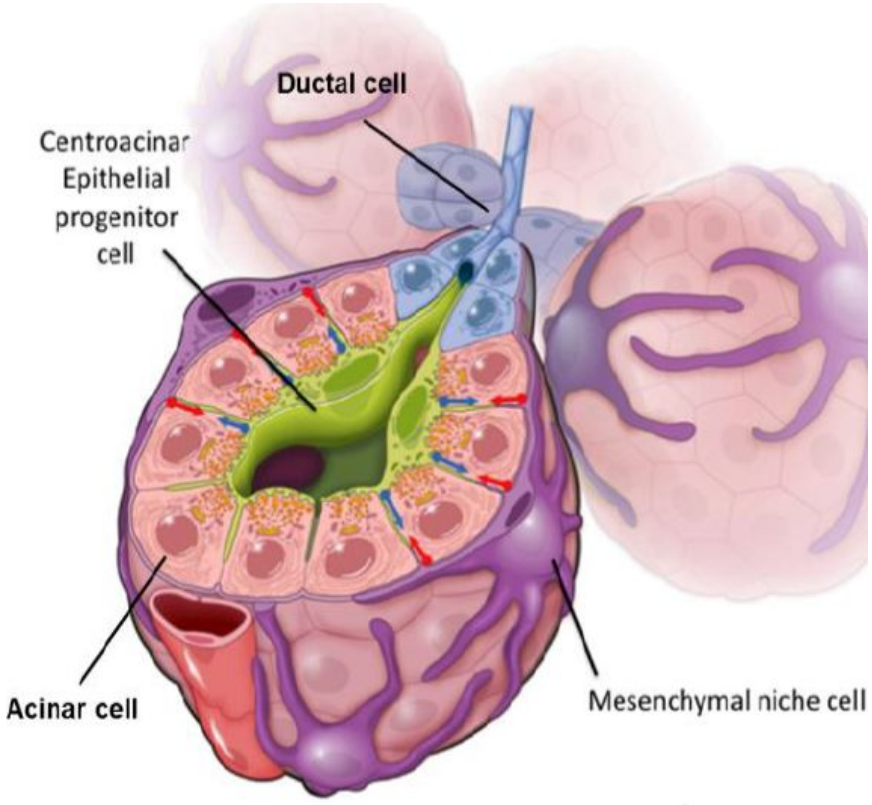


Figure 1. Schematic representation of the exocrine component of the pancreas. Acinar cells constitute the vast majority of the exocrine functional element and secrete digestive enzymes that are conveyed by the ductal cells into the small intestine. Together with mesenchymal and centroacinar cells, they conform the exocrine cellular composition. (Image is a kind gift of S. Leach, Darmouth Cancer Center, NH, USA).

PDAC is one of the deadliest cancers with a 5-year survival rate of 6%. One of the causes of this high mortality is its late diagnosis, consequence of the appearance of symptoms at advanced stages when the tumor has spread and become metastatic[4]. Unless there is substantial progress, PDAC will become the second most common cause of cancer-related deaths by 2030, due to its close relation with age, obesity and metabolic syndrome[5].

The precise cellular origin of PDAC still remains a core question. At the histological level, PDACs resemble ductal cells, displaying cuboidal shape, ductal antigen expression and growth into tubular structures[6]. GEMMs provide evidence that PDAC can arise from all exocrine epithelial cells[7,8,9,10,11]. Genetic lineage tracing studies have shown that ductal and endocrine cells have limited oncogenic capacity[10,12,13], while PDAC can develop efficiently from immature acinar cells that undergo a series of reprogramming processes known as ADM and pancreatic intraepithelial neoplasms (PanINs)[13]. It is therefore believed that the exocrine acinar cells of the pancreas are the predominant cells of origin for PDAC.

Loss of cell identity has been shown to be associated with tissue injury, representing a first step towards carcinogenesis, and it is believed that acinar cell differentiation can act as a tumoral suppressor mechanism in the pancreas[14,15,16,17]. Acinar cell identity is determined by specific gene programs controlled by well-defined DNA-binding TFs[1,18]. These gene programs are organized into transcriptional modules, set of genes co-regulated by the same transcription factor that binds specific DNA sequence motifs within cis-regulatory elements such as enhancers and promoters[19]. Precisely defining these modules and their degree of overlap is crucial for understanding the regulatory mechanisms of the biological functions responsible for the intricated relationship between loss of acinar identity, tissue damage and cancer.

Genome regulation depends not only on the linear sequence of the DNA, but also on its organization in a three-dimensional structure. Eukaryotic genomes are packed into nucleosomes, structures of ~146 bp of DNA wrapped around an octamer of histone proteins, which, in turn, are compacted to form the chromatin[20,21]. Highly condensed chromatin, known as heterochromatin, affects the accessibility of TFs and prevents the recruitment of RNA polymerase II to DNA in many contexts, resulting in the silencing of gene expression[22,23]. Although it has been shown that some of this compacted DNA can be transcribed, the mRNA is continuously turned over, avoiding translation[24,25,26]. On the contrary, the euchromatin is a less compacted state and regulatory elements such as enhancers and promoters are usually nucleosome-depleted regions where TFs can access and physically interact with the DNA to recruit the transcriptional machinery and promote the transcription of genes or, on the other hand, block the recruitment of RNA polymerase II to repress gene expression levels[27]. The modulation of chromatin structure and, therefore, the accessibility and activity of the underlying genes are highly influenced by a complex and dynamic regulatory network of DNA methylation events and chemical modifications of histone proteins that together constitute the epigenome[28,29].

Transcriptional regulation is a fundamental biological process which has been widely studied in order to better understand how gene expression levels are modulated[30]. TFs often interact with other TFs and co-factors to form complexes that bind to DNA and regulate the levels of transcription of different sets of target genes.

These complexes, as well as their exerted regulation, vary in different cell types and under different cellular conditions[31,32]. The ensemble of DNA binding events can be used to decipher the architecture of interactions between different TFs and target genes, constituting a transcriptional regulatory network[33,34].

To interpret the high degree of complexity characterizing regulatory mechanisms of biological systems, network biology has become a major strategy[35,36]. Through the construction of mathematical models based on qualitative and quantitative empirical measurements, it is possible to infer and reverse-engineer cellular functions by creating Gene Regulatory Networks (GRNs) representations[37,35].

The rapid development of molecular biology techniques and next-generation-sequencing (NGS) technologies, along with computational biology methods of analysis have opened the way for interrogating in a comprehensive manner the behaviour of complex systems, such as transcriptional regulation, through the construction of predictive network models by integrating multiple and complementary sources of data[38].

In this work, we present a comprehensive multi-omic approach to decipher the transcriptional network that governs acinar cell identity in mouse pancreas under homeostatic conditions. We integrate several layers of information from different NGS omic data using multiple computational biology methods to obtain robust results. Our approach starts from the most unbiased method to detect open chromatin regions (OCRs) in the pancreas (Figure 2A), followed by filtering (Figure 2B) and footprinting analysis (Figure 2C) restricted to active regulatory regions in acinar cells through the integration of different NGS methodologies (Figure 2D and E). This allows to identify specific transcription factor binding sites (TFBS) (Figure 2F) to build networks that model the genetic regulation underlying acinar cell identity. Ultimately, this resource can be useful to identify important transcriptional modules and their respective biological functions, providing a baseline model against which to compare subsequent examinations of GEMMs displaying different conditions such as PDAC. Moreover, the presented methodology can be applied to other organisms, such as human, and tissues to compare and cross-validate the obtained results.

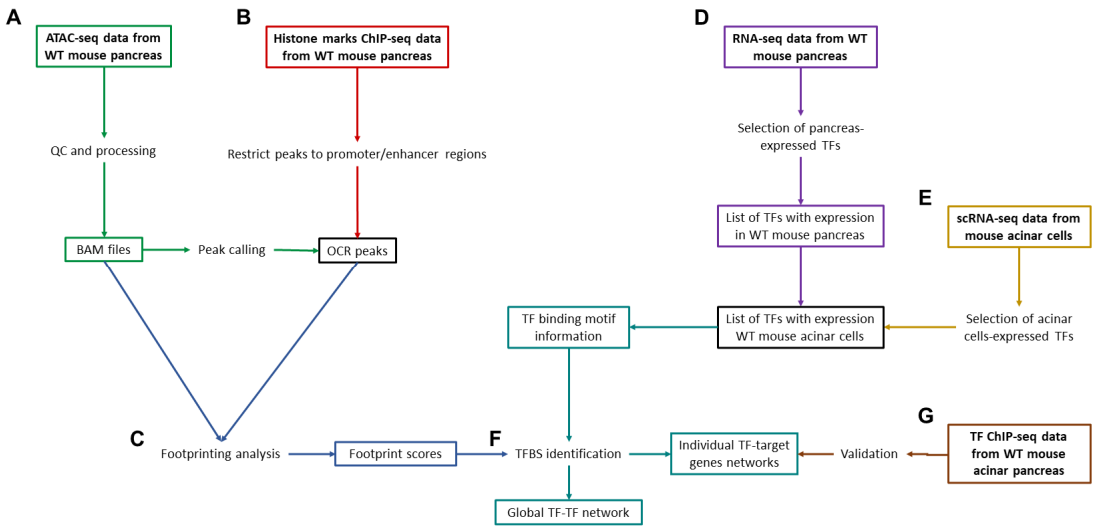


Figure 2. Project overview. A) Identification of OCRs in WT mouse pancreas by ATAC-seq data analysis. **B)** Integration of histone marks ChIP-seq data to filter and restrict the identified OCRs to promoter and enhancer locations as complementary approaches. **C)** Footprinting analysis on the different sets of OCRs to identify DNA protected sites from transposase cleavage due to protein binding. **D)** Selection of TFs expressed in WT mouse pancreas by integration of RNA-seq data. **E)** Integration of scRNA-seq data to filter the previously selected TFs, keeping only those expressed in pancreatic acinar cells. **F)** Integration of the footprint scores with TF binding motif information for the acinar-expressed TFs to identify specific TFBS and for construction of TF-TF interaction networks and individual TF-target genes networks. **G)** Validation of the networks built for individual TFs by comparison with ChIP-seq experiments performed for the same TFs.

Results

Identification of consensus open chromatin regions using ATAC-seq

Our first step in this work was to identify transcriptionally accessible regions, which represent potentially active regulatory regions and set up the basis for all the downstream analysis.

The Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq)[39] has become in recent years the gold standard technique to reveal open chromatin accessibility at a genome-wide level. It relies on the hyperactive Tn5 transposase, which inserts sequencing adapters into accessible chromatin regions[40]. The distribution of the Tn5 insertion signal defines the OCRs and allows for the detection of TF binding events occurring within these accessible regions. Bound sites are represented as a signal depletion because the Tn5 is not able to cut the DNA protected by protein binding. This lack of cleavage events inside OCRs is called footprint[41] (Figure 3) and can inform about TF binding distribution throughout the genome.

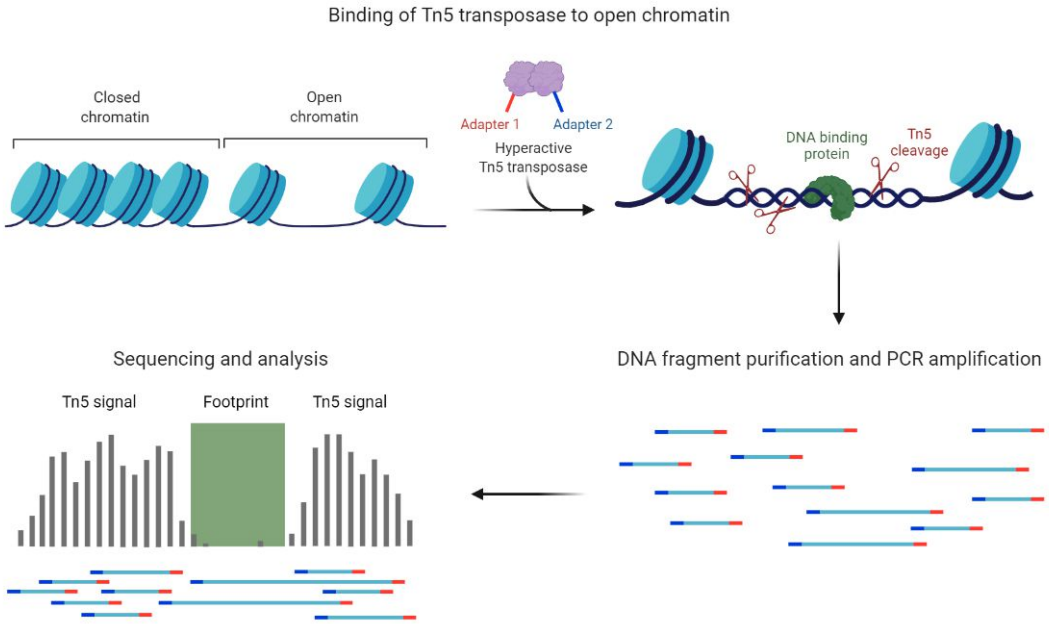


Figure 3. Footprint detection by ATAC-seq. The hyperactive Tn5 transposase accesses to the open chromatin regions to cleave the DNA and insert sequencing adapters. DNA regions occupied by proteins are protected from Tn5 cleavage and no insertions occur. DNA fragments are purified, amplified and sequenced to produce reads that can be mapped to the reference genome to generate signals corresponding to the Tn5 insertions into accessible regions. A depletion of the signal within these regions of high read coverage indicates the presence of protein binding events, also known as footprints.

We made use of publicly available ATAC-seq data from WT mouse pancreas to identify TF-binding events occurring in normal mouse pancreas. This data was

extracted from an ATAC-seq atlas consisting of 66 profiles from 20 different tissues[42] of both female and male mice. For the aim of this work, we focused on the 4 available pancreas profiles, corresponding to two adult female mice and two adult male mice.

The analysis of the raw data was performed using the ENCODE ATAC-seq pipeline developed by Anshul Kundaje's laboratory[43]. This pipeline allowed for an end-to-end quality control and processing from raw FASTQ files to peak calling between replicates. The obtained filtered Binary Alignment Map (BAM) files, containing the reads mapped to the reference genome, constitute the basal information for performing footprinting analysis into accessible regions in each of the replicates.

Female and male replicates were processed separately, resulting in two peak files containing the significant OCRs identified in each case. The consistency of the peak calls between replicates was assessed by the Irreproducible Discovery Rate (IDR)[44], discarding peaks with more than a 5% chance of being an irreproducible discovery (0.05 IDR).

In order to generate the most robust OCRs consensus set, we focused on the regulatory regions common to all four replicates. To do so, we merged both peak files making use of mergePeaks function from HOMER software[45] to obtain a consensus file where the peaks comprise the coordinates from both female and male replicates in the common regions. We obtained 38424 common peak coordinates between replicates that constitute a consistent representation of normal mouse pancreas active chromatin landscape and therefore were taken as the regions of interest to focus our subsequent analysis (Figure 4A).

Restriction of OCRs to specific functional regions by integrating histone marks information

Once determined the accessible chromatin regions by ATAC-seq analysis, we interrogated, separately, different functional regions of the active chromatin. To do so, we filtered the ATAC-seq identified OCRs by introducing information of epigenetic histone modifications. This approach also constitutes a cross-validation and a quality assessment of the previously defined OCRs.

Chromatin accessibility can be influenced by chemical modifications of the histone proteins, typically on their unstructured ends, which are often used as marks for transcriptional activation or repression. Moreover, some of them are known to be enriched in certain genomic locations such as enhancers or promoters. Therefore, we used this information to restrict the ATAC-seq based OCRs to perform complementary analyses focusing on specific regulatory regions.

We took advantage of ChIP-seq data from experiments performed in our laboratory for three different histone marks: H3K4me3, H3K27ac and H3K27me3. ChIP-seq is a method for genome-wide detection of protein binding to DNA. It combines the

use of specific antibodies targeting DNA binding proteins for chromatin immunoprecipitation with sequencing. In this case, it allowed us to determine the genomic coordinates where these histone marks were present in WT mouse pancreas. H3K4me3 and H3K27ac are epigenetic modifications associated with the activation of transcription, enriched in promoters[46] and both promoters and enhancers[47], respectively. On the other hand, the H3K27me3 mark is associated with the repression of nearby genes through the formation of heterochromatic regions[48].

ChIP-seq data analyses for these three histone marks were performed using RUbioSeq+, a command line multiplatform application that integrates automated and parallelized workflows for the analysis of new generation sequencing data[49]. The obtained peaks for the different replicates were intersected again using HOMER's mergePeaks function to obtain the consensus coordinates between all the replicates. We obtained 23053 consensus peaks between replicates for the H3K4me3, 41772 for the H3K27ac and 15173 for the H3K27me3 (Supplementary figure 1).

In order to restrict the OCRs identified by ATAC-seq to active promoter and enhancer regions, as well as to repressed regions as quality assessment, we intersected the sets of peaks identified for each histone mark with the ATAC-seq OCRs. In this case we used bedtools[50] intersect function to establish a specific overlap threshold and to keep in the output file the coordinates corresponding to the overlapping ATAC-seq peaks only.

In the case of the H3K27ac, as it is a mark for both promoters and enhancers, instead of intersecting the whole set of peaks, we divided it into two subsets, one corresponding to promoters, which also allowed us to compare and validate the results obtained with the H3K4me3 modification and another one enriched in enhancers. To do so, using bedtools, we intersected the H3K27ac consensus peaks with mouse transcription start sites (TSS) expanded 1 kb upstream and downstream to represent promoter regions. We defined as the H3K27ac promoter subset those peaks overlapping at least 1 bp with the TSS +/-1 kb regions, while the remaining were assigned to the enhancer-enriched subset.

After the intersection of the histone mark peak sets with the 38424 OCRs identified by the ATAC-seq analysis, we obtained 16408 OCRs overlapping with H3K4me3 mark (Figure 4B), 14788 OCRs that overlap with the peaks of the promoter subset of the H3K27ac modification (Figure 4C) and 10877 OCRs overlapping with the enhancer peaks subset of H3K27ac mark (Figure 4D).

As a quality control, we checked the consistency between the peaks identified by the H3K27ac promoter subset and the peaks marked as promoters by H3K4me3. We obtained 14348 common regions between both promoter peak files, which shows a good overlap and supports the quality of the ChIP-seq data, as well as the accuracy of our splitting of the H3K27ac peaks (Supplementary figure 2A).

In order to identify functional elements in the whole OCRs identified by ATAC-seq and in the different sets of regions filtered by histone marks ChIP-seq data, we annotated the genomic locations of the peaks by genomic feature association analysis with HOMER. We observed that in the case of the H3K4me3 mark and the promoter subset of the H3K27ac more than half of them corresponded to TSS or exons, of which around 85% were first exons, which are close to promoter regions. In the case of the enhancer subset of the H3K27ac, we found that almost half of the peaks corresponded to intergenic regions and this set was depleted of promoter regions (Figure 4E).

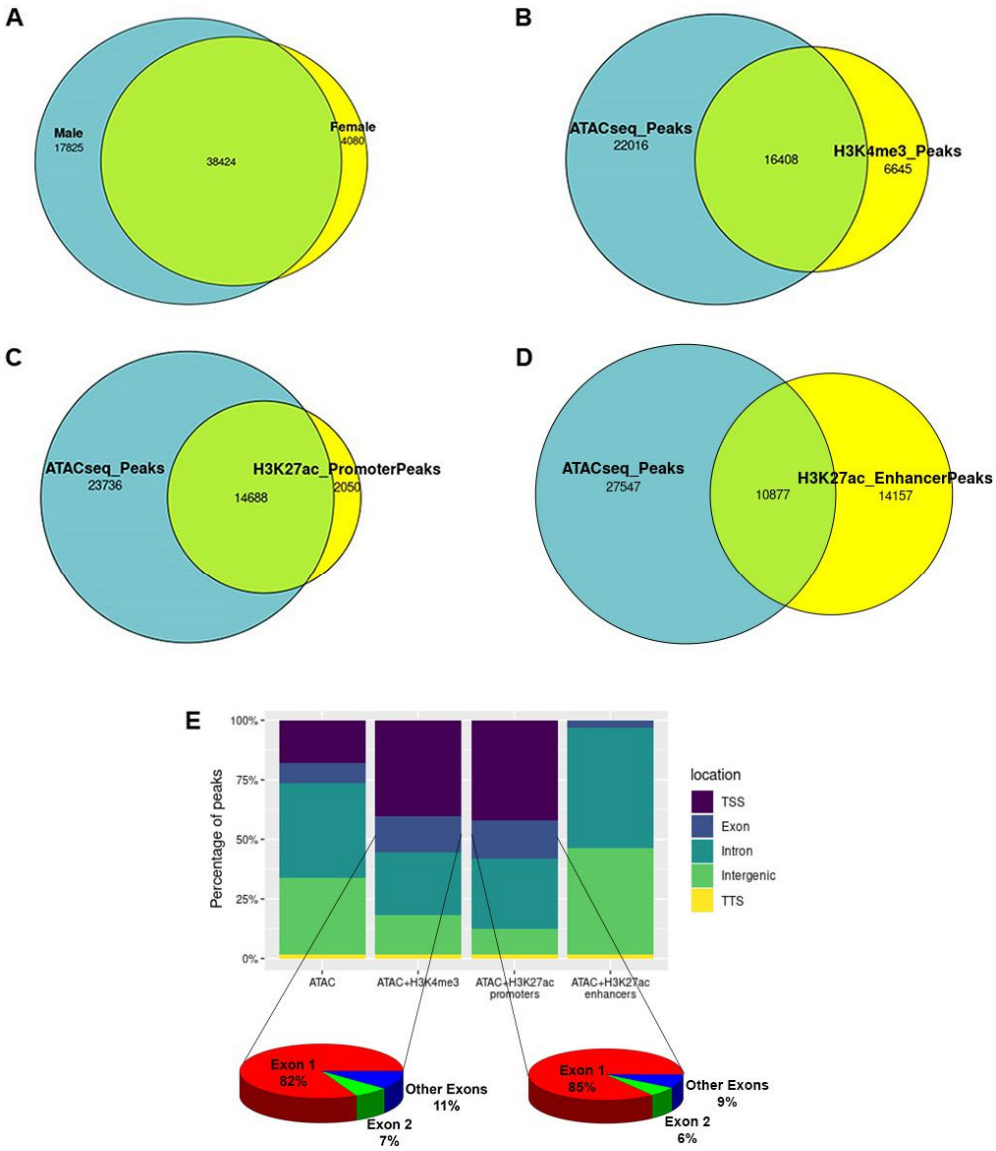


Figure 4. Sets of open chromatin regions used for footprinting analysis. A) OCRs identified by ATAC-seq in WT mouse pancreas. There are 38424 identified consensus regions where the Tn5 cut in both female and male replicates. **B)** Overlap between the consensus OCRs identified by ATAC-seq and the regions marked as active promoters by H3K4me3 epigenetic modification. 16408 overlapping peaks were identified. **C)** Overlap between the identified OCRs and the regions marked as active promoters by H3K27ac (promoter subset) epigenetic modification. 14688 overlapping peaks were identified. **D)** Overlap between the consensus OCRs and the regions marked as active enhancers by H3K27ac (enhancer subset) epigenetic modification. 10877 overlapping peaks were identified. **E)** Genomic location distribution of the OCRs as a whole and filtered by different histone marks.

As negative control, we assessed the overlap between the peaks of H3K27me3, which is a mark associated with transcriptionally repressed chromatin, and the ATAC-seq identified OCRs. As expected, only a small fraction (946 out of 38424 OCRs) overlapped with peaks marked by H3K27me3 as repressed regions (Supplementary figure 2B).

In summary, the integration with histone marks showed that OCRs identified by ATAC-seq were consistently enriched in transcriptionally active chromatin. Furthermore, this integrative approach allowed us to create different OCRs subsets that will facilitate downstream analyses to separately interrogate different functional regions of the active chromatin in WT mouse pancreas.

Identification of TFBS through footprinting analysis

In order to determine specific TF binding events to DNA in the previously generated OCRs subsets, we further examined the Tn5 cut signal through footprinting analysis to identify signal depletion events within accessible regions due to the presence of DNA bound proteins. We made use of TOBIAS, a collection of command-line bioinformatic tools specifically developed for this purpose[51]. We also restricted the analysis specifically to active regulatory regions in acinar cells through the integration of other NGS technologies using different computational biology methods. These computational methods allowed us to measure the TF expression levels in the different pancreatic cell types in order to filter the identified TFBS, excluding information related to non-acinar-expressed TFs.

The first thing to consider when performing footprinting analysis on ATAC-seq data is that the Tn5 transposase, like the DNaseI enzyme used in DNase-seq chromatin accessibility assay, has preference for specific DNA sequences[52,53]. This causes an intrinsic sequence-dependent transposition site bias that interferes with the identification of footprints[54,55]. Therefore, as a first step, it was necessary to correct that bias for an accurate footprint prediction. To do so, we used the ATACorrect TOBIAS module, which takes the mapped ATAC-seq reads and the peak file with the regions of interest and, using a Dinucleotide Weight Matrix (DWM)[56], calculates an expected Tn5 insertion signal for each genomic region. This signal corresponds to the background cleavage bias of the transposase, which is subtracted from the observed signal to yield a corrected signal. That correction allowed to identify regions with weaker cut signal than expected, suggesting DNA protection from Tn5 cleavage due to the presence of protein binding (Figure 5A).

In order to identify potential footprints, we quantitatively assessed the corrected cut signal by calculating footprint scores across regions using the ScoreBigwig tool. These scores are calculated as the difference of the background mean signal and the footprint mean signal, thereby taking into account not only the

depletion of signal, but also the accessibility of the flanking regions, which improves the prediction of bound TFs with weak footprints (Figure 5B).

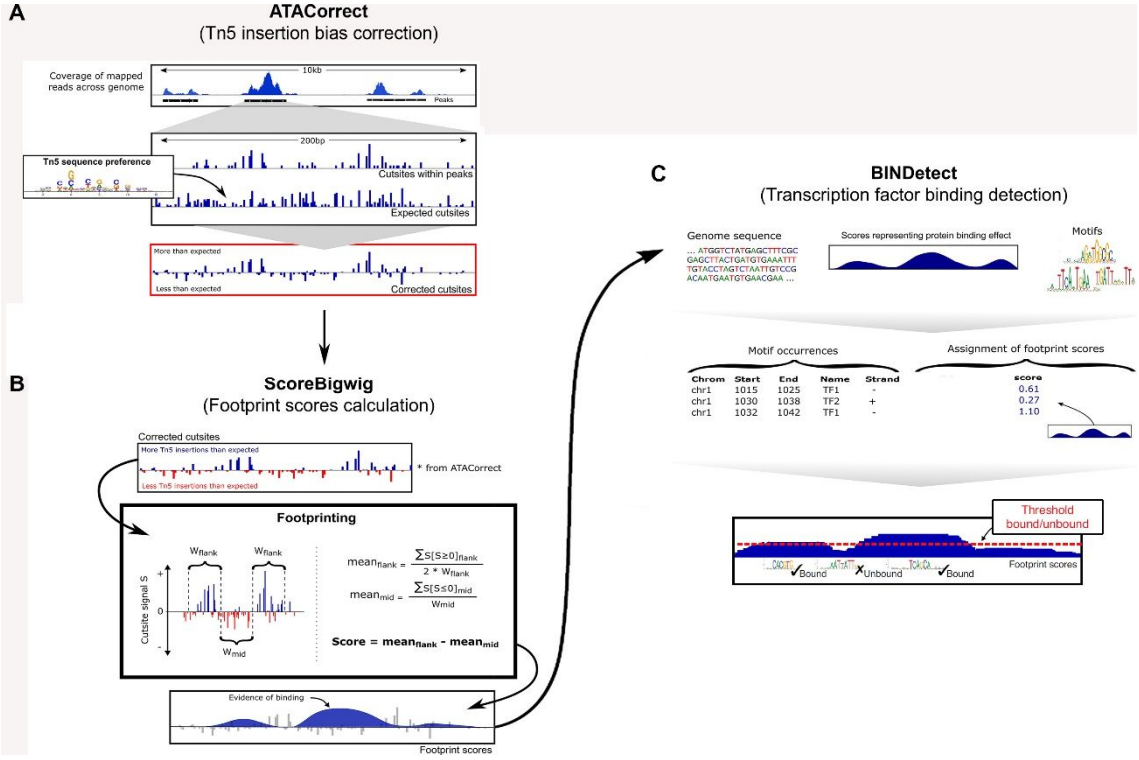


Figure 5. TOBIAS footprinting analysis. A) The ATACorrect module estimates the Tn5 sequence preference based on the observed reads and calculates the expected Tn5 insertion signal for each genomic region. Then, this biased signal is subtracted to obtain the corrected signal. **B)** The ScoreBigwig module performs footprinting to estimate the presence of transcription factor binding events. This tool calculates a footprint score based on both the depletion of insertion signal and the accessibility of the regions flanking the footprint. **C)** The BINDetect module integrates the footprint scores with information of transcription factor binding motifs to estimate the specific binding sites across the genome. It also establishes a threshold based on the footprint scores to determine the bound or unbound status of the identified sites. Scheme modified from Bentsen, M. *et al.*, 2020.

Once we had quantitatively predicted DNA footprints across regions, we integrated the calculated scores with TF binding motif information to estimate the specific binding coordinates of individual TFs. Since our work was focused on the regulation of acinar cell identity, we integrated data from other computational biology methods to restrict the introduced motif information to acinar-expressed TFs.

Firstly, we analysed RNA-seq data from WT mouse pancreas coming from experiments performed in our laboratory and publicly available. This data allowed us to measure and rank the expression levels of TFs in WT mouse pancreas. A threshold of 1 RPKM was set to differentiate between expressed and non-expressed TF in normal mouse pancreas, as expression values below 1 RPKM are considered to represent noise rather than true biological signal. In supplementary table 1, we show the top 100 TFs ranked by measured expression levels in WT mouse pancreas.

The representation of the distribution of TFs expression levels showed that, surprisingly, only a few of them are highly expressed, while the rest of TFs compose a long queue with lower expression levels. Interestingly, a similar distribution was observed in the case of human pancreas data, coming from the Genotype-Tissue Expression (GTEx) project[57]. In both cases, Xbp1 had the highest expression levels, followed by Rbpjl, Bhlha15 (also known as Mist1) and Atf4 and a long queue of gradually less expressed TFs. This similarity between the TFs expression levels in both mouse and human pancreas indicates that the results of this project could, to some extent, be extrapolated to human data (Figure 6).

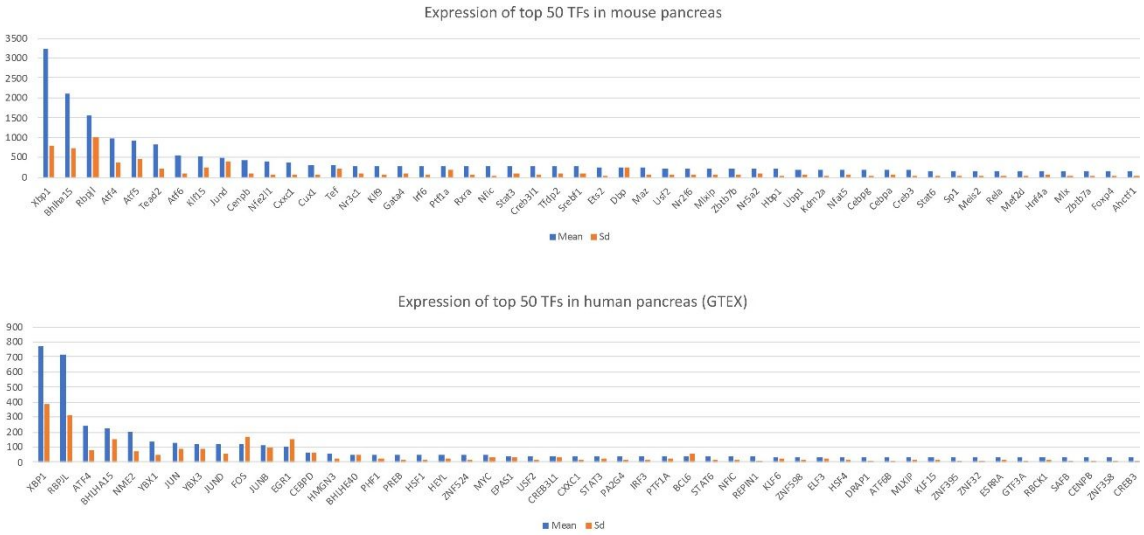


Figure 6. Mean expression and standard deviation comparison between the top 50 expressed TFs in mouse and human pancreas. Similar mean expression distribution is observed for mouse and human data. Xbp1 is highly expressed compared to the overall expression levels and is followed by Rbpjl, Bhlha15 and Atf4 among the top 5 most expressed TFs in both human and mouse data. A long queue of gradually less expressed TFs is observable in both cases.

The measured expression mostly corresponds to acinar expression, as this cell type constitutes around 85% of pancreatic mass. However, since the scope of our work was the transcriptional regulation of acinar identity, we introduced acinar cell expression data from scRNA-seq to discard TF binding motif information from TFs expressed specifically in non-acinar pancreatic cells. In contrast to bulk RNA-seq, scRNA-seq method can capture the transcriptome of individual cells, allowing to assess the biological properties of specific cell types. We took advantage of mouse pancreas scRNAseq data analysed in our group consisting of WT acinar cells, acinar cells treated with cerulein and OSKM reprogrammed acinar cells (Red circles in figure 7B). To be conservative, we considered as acinar-expressed those TFs with expression in ≥ 1 acinar cell in the whole dataset (Figure 7A). The final list with the selected acinar-expressed TFs is shown in supplementary table 2.

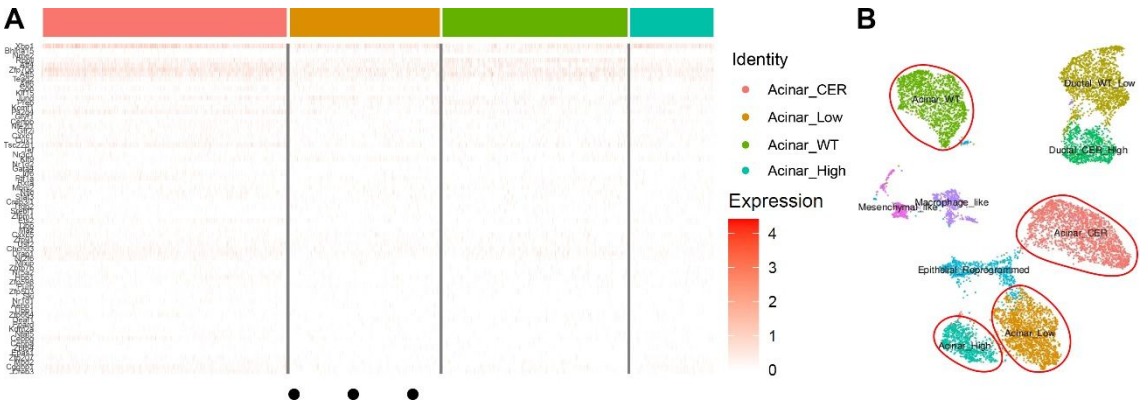


Figure 7. Identification of transcription factors expressed in acinar cells by single cell RNA-seq. A) Heatmap showing the expression of each TF in the different sets of acinar cells assessed. **B)** UMAP plot showing the clusters of different cell types in the scRNA-seq, with the acinar cells surrounded by a red circle. TFs with expression signal in any cell from wild type acinar cells (green), acinar cells treated with cerulein (pink) or OSKM reprogrammed acinar cells (orange and blue) were considered as acinar-expressed TFs for downstream analysis.

The previously calculated footprint scores across regions were linked with Position Frequency Matrices (PFMs) for the selected acinar expressed TFs to identify specific TFBS. These PFMs, consisting of nucleotide counts per position representing the binding motifs of each TF, were extracted from CIS-BP and JASPAR databases. The association between these motifs and footprint scores was assessed by using the BINDetect module from TOBIAS, which allowed us to set a footprint threshold to distinguish between bound and unbound TFBS (Figure 5C).

In order to visualize the different shapes and patterns of footprint signals between bound and unbound sites, we created aggregated views of these signals across regions making use of TOBIAS PlotAggregate function. This allowed us to analyse in more detail the footprint signals for specific TFs and to compare the different shapes between pairs of TFs commonly working as co-regulators. As example, in figure 8 we show the aggregated footprint signals obtained in one of the female replicates for Nfic and Nr5a2, two acinar-expressed TFs of particular interest in our group due to their role in the maintenance of acinar identity in a context of inflammatory response of the pancreas[58]. While Nfic shows a canonical footprint signal, with a clear depletion between two peaks indicating chromatin accessibility, Nr5a2 has a weaker and more irregular signal. Although the shown aggregated footprint signals for Nr5a2 do not present a canonical shape, the represented regions passed the threshold to be identified as sites bound by Nr5a2 since the accessibility of the flanking regions is also taken into account to determine the bound or unbound state of TFs.

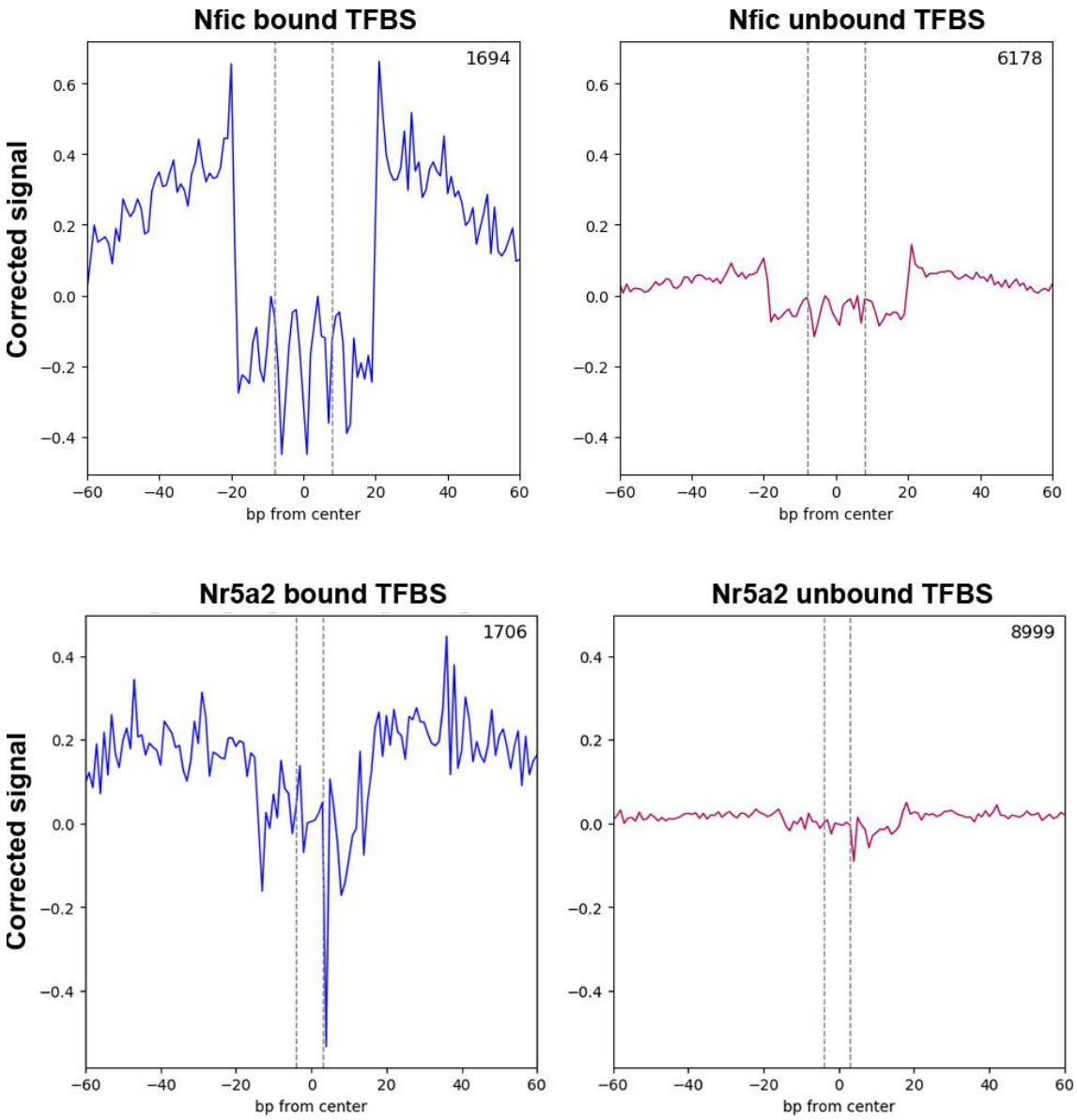


Figure 8. Aggregated footprint signals for bound/unbound sites. Aggregated plots showing the Tn5 cut signal across the identified OCRs containing the Nfic (top) and Nr5a2 (bottom) binding motifs. On the left, mean signals across regions that passed the footprint threshold to be classified as bound TFBS. On the right, aggregated signals across regions where the corresponding transcription factor binding motif was identified but did not pass the footprint threshold. The plots are centered around the identified TF binding motif, delimited by a dashed line.

Altogether, we integrated TF binding motif information for the selected acinar-expressed TFs with the previously calculated footprint scores to estimate specific TFBS across the OCRs identified by ATAC-seq and across the different subsets of regions filtered by histone marks. The resultant TF-annotated footprints maps constituted the stepping stone for the construction of regulatory networks in different functional regions of the active chromatin in WT mouse pancreas.

Transcriptional regulatory networks construction

Once we had identified specific binding sites for individual TFs, we associated the estimated TFBS with the corresponding target genes to elucidate the regulation exerted by each TF. To do so, we used HOMER annotation function to assign each TFBS to the gene corresponding to the nearest TSS. Then, taking the identified TF-gene pairs and using the CreateNetwork tool, we could model the interactions between different TFs and their target genes to represent regulatory networks underlying acinar identity in different functional regions.

A gene regulatory network is a mathematical representation of interactions between TFs and genes, represented as nodes, which are connected by directed edges indicating the regulatory relationship between them. We started building TF-TF networks, excluding genes not coding for TFs, to obtain a global view of the interactions of TFs occurring in pancreatic acinar cells. We also excluded the zinc-finger proteins (ZFP) because their functions are mostly unknown, which could interfere with the interpretation of the resulting networks and therefore should be explored separately. Furthermore, most of ZFP present poorly defined motifs, which could lead to an inaccurate representation of this kind of TFs in the networks.

Networks modelling TF-TF interactions were built for the whole set of OCRs identified by ATAC-seq, for the OCRs restricted to promoters by H3K4me3 filtering, for the OCRs filtered by the promoter subset of H3K27ac and for the OCRs filtered by the enhancer subset of H3K27ac. Four networks were built in each case, corresponding to the four initial replicates, and were visualized with Cytoscape[59]. To obtain a robust representation of the TF interactions, we intersected the replicates using Cytoscape's merge tool to generate a consensus network for every approach.

The obtained TF-TF networks consist of 336 nodes representing the acinar-expressed TFs and 8637 edges in the ATAC-seq network (Figure 9A), 4556 edges in the network restricted to promoters by H3K4me3 (Supplementary figure 3A) and 3968 edges in the network restricted to promoters by H3K27ac (Supplementary figure 4A). The network restricted to enhancers by H3K27ac consists of 331 nodes and 2578 edges (Supplementary figure 5A). These representations gave us a global view of the great number of interactions occurring between TFs expressed in acinar cells and allowed us to focus on specific regions by interrogating the global networks.

Important nodes playing a central role in the network could be identified taking advantage of NetworkAnalyzer module from Cytoscape. This tool allowed us to analyse the topological properties of the different networks based on local and global topological methods. A local method measures the relevance of a node in the network by considering the relationship between the node and its direct neighbours, whereas a global method considers the relationship between the node and the entire network. These analyses allowed us to rank the relevance of the TFs in the network at two

different scales. Specifically, we used two metrics, the degree, a local method that measures the number of nodes directly interacting with the node being assessed; and the betweenness centrality, a global method that measures the centrality of a node based on the number of shortest paths. Thus, considering that between each pair of nodes in the network exists at least one shortest path, the betweenness of a node is calculated by the summatory of the fractions of shortest paths between each pair of nodes that pass through the node being assessed.

A strong correlation was found between the degree and the betweenness centrality scores of the nodes in all cases (Figure 9C and supplementary figures 3C, 4C and 5C). This similarity between both local and global methods when ranking the nodes of the network allowed us to determine the relevance of each node in the network in a consistent way. Therefore, we could build subnetworks to focus on the most important nodes and their interactions for further analysis.

Global TF-TF networks were ranked by degree to generate subnetworks for the top 20 TFs based on this metric, obtaining consistent results between the ATAC network (Figure 9B) and both H3K4me3 (Supplementary figure 3B) and H3K27ac promoter (Supplementary figure 4B) networks. Klf (Klf4, Klf5, Klf11, Klf15) and Sp (Sp1, Sp2, Sp4) family members were observed as the most relevant nodes, but also other key transcription factors involved in the maintenance of a healthy acinar cell differentiated state like Bhlha15 and Nr5a2[60,61]. On the other hand, Klf or Sp members were not found among the most relevant nodes in the enhancer network, suggesting that their regulatory activity is restricted to promoters. We found Gata6 and Gata4, two important TFs in this context due to their role as regulators of epithelial differentiation in pancreas and their controversial function in PDAC[62,63]. Again, we found Bhlha15 among other highly expressed TFs in pancreas (Supplementary figure 5B). Considering the strong correlation between degree and betweenness methods when ranking nodes, we focused on the former to build these subnetworks and to perform further analysis.

The correlation between the ranking of nodes based on topological measurements and the ranking of TFs based on expression levels was assessed starting from the assumption that TFs playing a central role in the network should be more expressed than others. A global network with all TFs and target genes interactions was also generated for more realistic comparison. A weak correlation was found for the TF-TF networks (Figure 9D and supplementary figures 3D, 4D and 5D) and no correlation was observed for the global network (Supplementary figure 6).

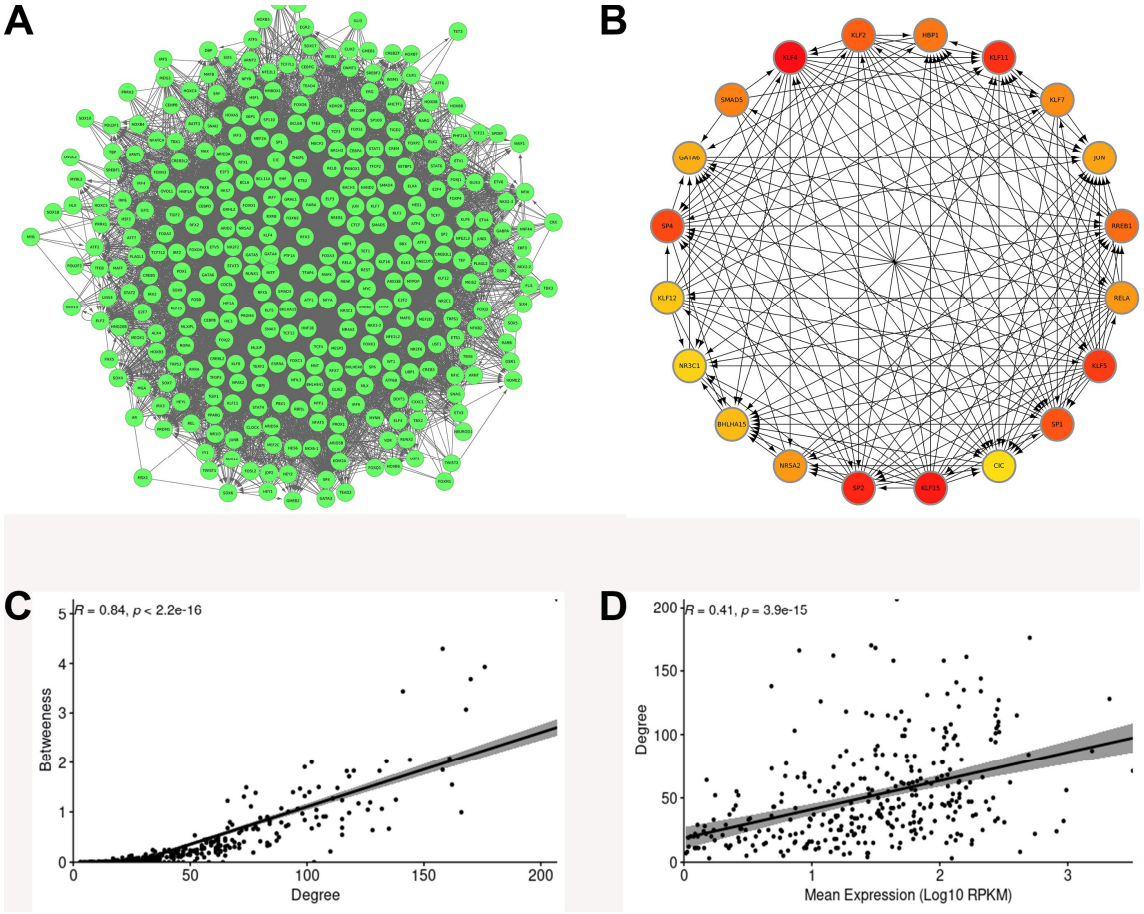


Figure 9. Construction and topological analysis of TF-TF network (ATAC-seq peak set). A) Global view of the TF-TF network for the whole ATAC-seq set of peaks. **B)** Top 20 nodes (TFs) of the network ranked by degree. **C)** Correlation between degree and betweenness centrality measures for the ATAC network. A strong correlation was found between both local and global methods. **D)** Correlation between nodes degree and expression levels of TFs. A weak correlation was found. Also see supplementary figures 3-5.

In addition, we followed an alternative candidate approach to generate networks consisting of individual TFs and their target genes to study in more detail the regulation of biological functions through the identification of co-regulatory events. We started building networks for Gata4, Gata6, Foxa2, Nr5a2, Nfic, Rbpjl, Ptf1a, Bhlha15 and Hnf1a, TFs of interest for our group, with which we have worked due to their important role in different regulatory functions in acinar pancreas. On the left panels of figure 10, we show a pairwise comparison of the regulated genes by each TF based on the individual networks. The number of target genes identified individually for each TF is indicated on the heatmap axis, between brackets. The number of common target genes regulated by each pair of TFs is indicated on the heatmap cells. As expected, Gata4 and Gata6 presented the highest number of co-occurrences in both promoters and enhancers, since both are involved in the regulation of similar biological functions in the pancreas, often working as co-regulators[64,65,63,66,67]. It is also remarkable the lower number of target genes for Hnf1a in the enhancer network, as well as the increased number of regulated genes by Rbpjl and Nfic in the same network, which gives an idea of the proximal and distal regulation exerted by each TF.

On the right panels of figure 10, we generated subnetworks with the top 50 nodes comparing the studied individual TF networks ranked by degree. The assessed TFs were displayed on the periphery, with the most relevant target genes on the centre. Among these target genes we obtained important acinar enzymes such as Cela1, Cela3b, RNaseI (Figure 10A-C), Amy2b and Pnlip (Figure 10A). We also obtained TFs involved in the development of PDAC such as Foxo3 (Figure 10A), Jun, Onecut1 (Figure 10B) and RREB (Figure 10C) [68,69,70,71,72].

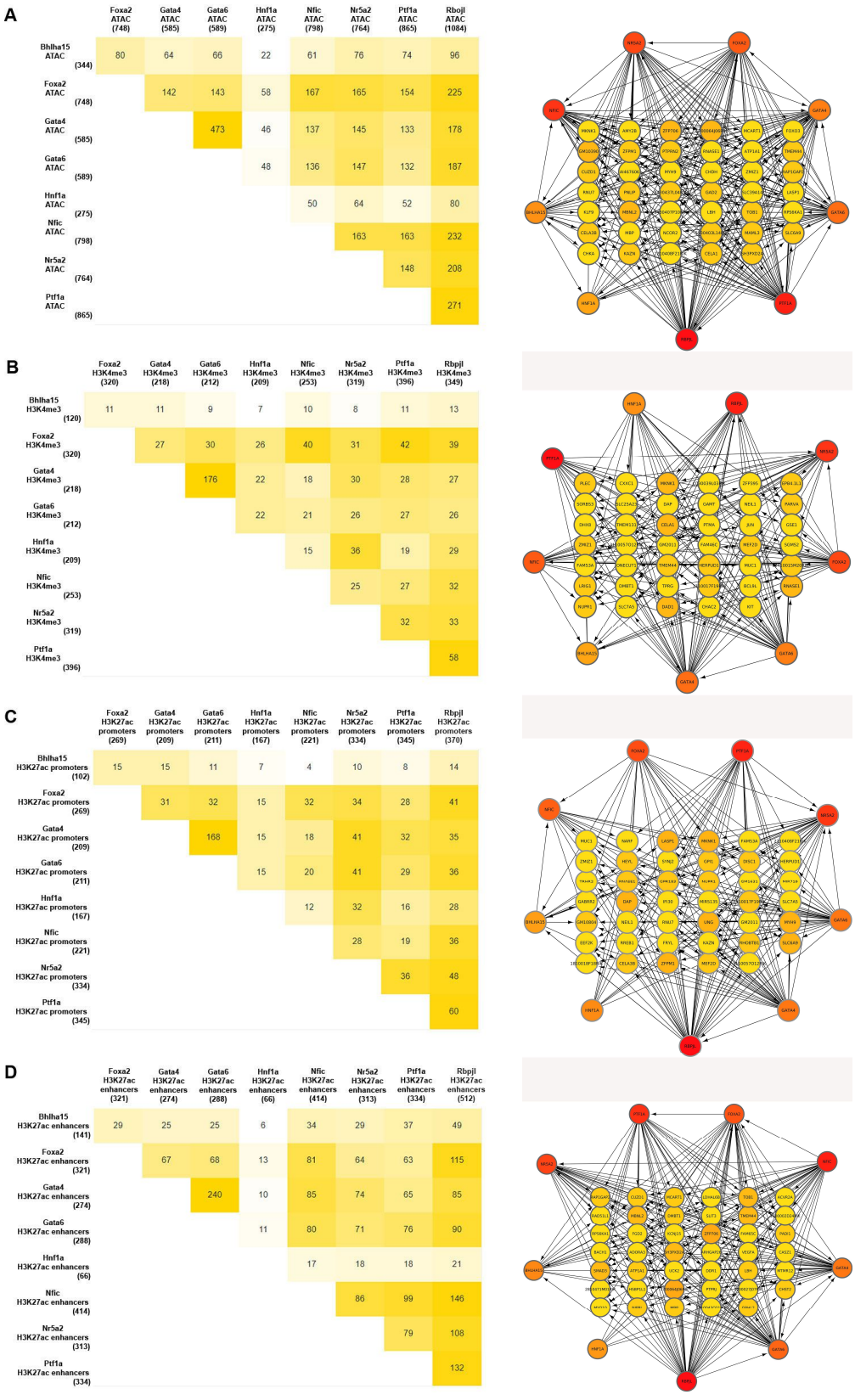


Figure 10. Construction and topological analysis of individual TF-target genes. A) ATAC-seq peaks-based networks. **B)** Networks based on OCRs restricted to H3K4me3 peaks. **C)** Networks based on OCRs restricted to H3K27ac promoter peaks subset. **D)** Networks based on OCRs restricted to H3K27ac enhancer peaks subset. On the left, pairwise comparison of the target genes identified for each TF. On the right, network for the top 50 nodes ranked by degree from the merge of all individual TF networks. Regulated genes are located on the centre, while the evaluated TFs are on the periphery.

In addition, we took advantage of ChIP-seq experiments performed for these TFs for which we built individual TF networks to validate our approach. To do so, we annotated the ChIP-seq coordinates to determine the corresponding target genes of each TF. For both network and ChIP-seq, we filtered the identified genes to keep only those with expression in pancreas. As we did before, we integrated RNA-seq data from WT mouse pancreas to rank the gene expression levels and we kept those with at least 1 RPKM.

We also introduced motif information to check which of the peaks identified by ChIP-seq were associated to their canonical binding motif for further analysis. In each case, the motif was identified by performing a *de novo* motif analysis with HOMER software on the ChIP-seq identified coordinates, obtaining the Position Probability Matrix (PPM) for the most enriched motif. These PPMs consist of a normalization of PFMs, showing nucleotide probabilities per position to represent TF binding motifs.

Therefore, to assess the consistency of the results, we compared the identified target genes from the networks with those from the ChIP-seq data. Moreover, for a more comprehensive analysis, we divided the ChIP-seq peaks into two subsets based on their association, or not, with the corresponding *de novo* identified motif. The annotated target genes from each subset were compared again with the target genes identified from the networks to assess the consistency of the results.

In figure 11, we represent the percentage of genes identified in the individual TF networks that were also identified by ChIP-seq analysis for the same TF. As shown in the heatmaps, the target genes identified in the individual networks built based on the whole ATAC-seq OCRs, as well as in the networks restricted to promoter and enhancer regions, were consistent with the target genes identified in the ChIP-seq for Gata4, Gata6, Foxa2, Nr5a2, Nfic and Rbpjl. This consistency is especially high in the enhancer networks, also for Ptf1a.

In general, many of the identified target genes were lost when restricting the ChIP-seq peaks with motif information, although the results for Gata4, Gata6 and Nr5a2 remained very consistent. Surprisingly, in the case of Foxa2 and to a lesser extent in the case of Rbpjl, the consistency of the results was higher for the ChIP-seq peaks not matching the motif compared with the peaks matching the motif. This suggests a discrepancy between the motifs used in network and ChIP-seq approaches to identify specific TFBS and a low accuracy of the *de novo* canonical motif identified from the ChIP-seq peaks, which was used for the motif/no motif split.

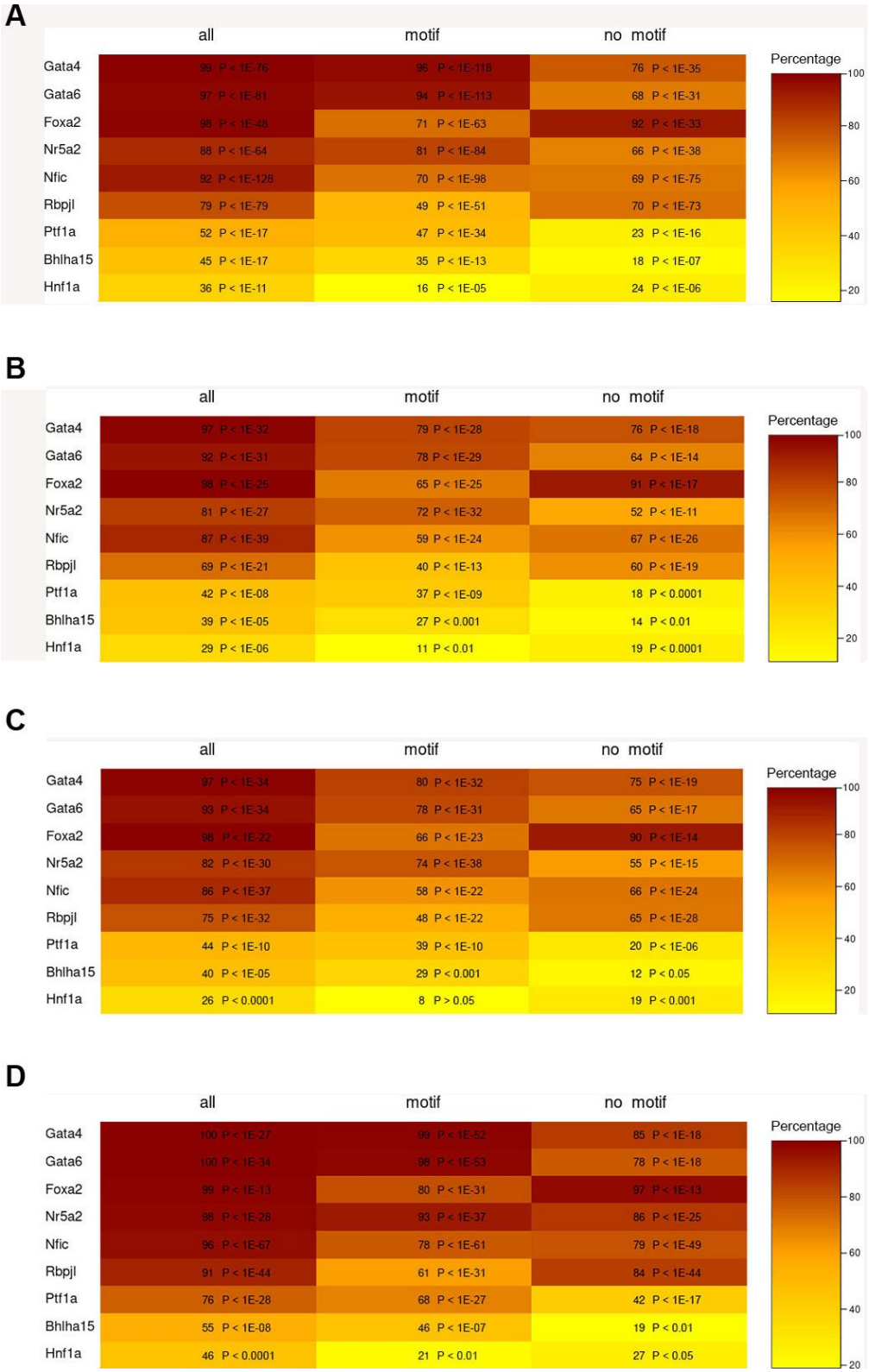


Figure 11. Validation of the results of individual TF-target genes networks. A) ATAC-seq peaks-based networks. **B)** Networks based on OCRs restricted to H3K4me3 peaks. **C)** Networks based on OCRs restricted to H3K27ac promoter peaks subset. **D)** Networks based on OCRs restricted to H3K27ac enhancer peaks subset. Numbers and colours of each cell represent the percentage of target genes identified by our multi-omic approach in each of the individual networks that were also identified by ChIP-seq analysis for the corresponding TF. In each heatmap, from left to right, validation of network results with all the ChIP-seq identified TFBS, with the regions associated to the corresponding TF binding motif and with the regions not matching the TF binding motif. P-values are indicated in each case.

In order to assess the significance of the common results between the network and ChIP-seq approaches, we generated random permutations of the ChIP-seq regions through the reference genome. Using bedtools shuffle function, each peak was repositioned on a random chromosome at a random position, while preserving the original size and strand. The permuted regions of each TF ChIP-seq were annotated to determine the target genes based on the nearest TSS and the results were compared with the target genes obtained in the corresponding individual TF network. Next, to determine the statistical significance of the results, the number of common genes identified between the individual TF network and the permuted regions from the ChIP-seq analysis of the same TF was compared with the number of common genes identified previously from both the network and the original non-permuted ChIP-seq peak regions. The common results between all the assessed individual TF networks in the different functional regions and the corresponding ChIP-seq data showed strong statistical significance (Figure 11). These statistical analyses validate the construction of individual networks as a useful starting point to infer the genes regulated by each TF expressed in pancreatic acinar cells.

Discussion

In this project, we have analysed and integrated different NGS omic data using multiple computational biology methods to generate networks that model the regulation underlying acinar cell identity in WT mouse pancreas. Starting from ATAC-seq data to detect OCRs, we have integrated histone marks ChIP-seq data to perform parallel approaches restricted to specific functional regions in order to distinguish between proximal and distal DNA regulatory events in promoters and enhancers, respectively. We have also integrated RNA-seq and scRNA-seq data to restrict the analysis to TFs specifically expressing in pancreatic acinar cells that, together with footprinting analysis, allowed us to identify specific TFBS. Based on this information, we generated in a first place, TF-TF networks modelling the hierarchical interactions between TFs involved in the regulation of acinar cell identity in normal mouse pancreas. On the other hand, we built subnetworks for individual TFs and their target genes for detailed analysis of their regulatory activity.

Degree and betweenness-based analyses were performed on the generated TF-TF networks to rank the nodes involved in the regulatory network. Degree and betweenness centrality constitute two topological measures frequently used to characterize the importance of nodes in a network[73,74,75,76,77]. Considering the high correlation obtained between both methods, we focused on the degree score, which showed consistent results between the network based on the whole ATAC-seq OCRs and the two networks restricted to promoter regions by H3K4me3 and H3K27ac. As shown in figure 9B and supplementary figures 3B and 4B, Klf and Sp family members emerged as the most relevant nodes in the networks following topological measures, which are known to be involved in regulation of growth, proliferation and migration of pancreatic cancer cells[78,79,80,81,82,83,84]. However, these families of TFs present GC-enriched binding motifs, which could lead to their overrepresentation due to a non-specific association of the TFs to non-related GC-rich functional elements. Similarly, there could be a wrong association between TFs and motif sequences belonging to other members of the same family, as they present very similar TF binding motifs and, therefore, should be taken into consideration when interpreting the results.

Based on these topological analyses of TF-TF networks, we can select important nodes and build individual networks for the selected TFs and their target genes for a detailed analysis of the regulation exerted by each TF.

The results obtained from the assessed individual TF-target genes networks allowed us to perform pairwise comparisons of the targets obtained for each TF to identify co-regulation events. The highest number of common target genes was observed for the Gata4 and Gata6 comparison. On the one hand, these results are expected from a biological perspective, as both Gata4 and Gata6 play an important role in mouse pancreas organogenesis, regulating similar and complementary cell

functions, and have gained especial attention in the past years due to their controversial role in PDAC and patient outcome[62,63,64]. Nevertheless, as well as with the results obtained from the topologically ranked TF-TF networks, we should consider the high similarity between both TF binding motifs, which could lead to false positive matches between one TF and the partner motif. Therefore, in an alternative and more conservative approach we could integrate both results, focusing on TF families instead of specific TFs with highly similar TF binding motifs.

To better understand this co-regulation, we compared the Gata4 and Gata6 target genes identified in promoter networks with the annotated genes in TSS regions from ChIP-seq experiments for the same TFs. We also performed this comparison for Nfic and Nr5a2, which unlike Gata4 and Gata6, present very different binding motifs, but also regulate similar functions in the pancreas, especially interesting in the maintenance of acinar cell differentiated state in the context of inflammatory response[58]. The results of these comparisons showed that the identified common target genes between the promoter networks of Gata4 and Gata6, as well as the common targets between Nfic and Nr5a2 promoter networks, were consistently identified as common between the corresponding TF ChIP-seq analyses. However, some of the target genes obtained as specifically identified by one of the networks were also identified by both TF ChIP-seq analyses. Taking into account that the number of target genes identified by the networks is lower than the number of genes identified by ChIP-seq, these results suggest that our approach could be too restrictive when identifying specific TFBS, leading to subnetworks that represent only a part of the real interaction landscape. Therefore, it would be interesting to relax some of the thresholds applied in each analysis and assess the consistency of the results. An improvement in the detection of common target genes between the network and the ChIP-seq analyses, while maintaining the specificity, would indicate that we can be more permissive in regard to the applied thresholds.

The methodology described in this work can be a useful approach to interrogate the regulatory network underlying the identity of pancreatic acinar cells in WT mouse. However, as mentioned, there are some caveats that should be considered to optimize this methodology and the scope of the results.

- Since TFs belonging to the same family present similar binding motifs, this could lead to miss association between TFs and binding motifs of other TFs of the same family. Therefore, instead of describing separately the regulation exerted by TFs with highly similar binding motifs, focusing on families of TFs could avoid false positive results.

- Another consideration following the previous constraint is that there are families of TFs with GC-rich binding motifs, as Klf and Sp families, which emerged in our topological analysis as the most important TFs. Nevertheless,

the presence of isochores (large GC-rich DNA regions) throughout the genome could lead to non-specific associations of these TFs with other non-related GC-rich DNA regions and therefore, to their overrepresentation. To improve the specificity on the association of footprints with TFs exhibiting unspecific sequence motifs, our future work will be focused on introducing new layers of information such as TF binding motif energy measures to strengthen the identification of specific TFBS[85,86,87].

- Regarding the target genes identified by our networks, we have observed that our approach represents only a fraction of the real regulatory landscape compared with the homologous ChIP-seq analysis. Therefore, future efforts should be centered on further studying the parameters and thresholds of the different computational methodologies applied in this work in order to find a good balance between scope and precision on the detection of TFBS. Our next step will be to relax the restrictiveness on the selection of OCRs coming from ATAC-seq, making use of all the significant identified regions (filtered by 0.05 IDR) from both male and female replicates instead of selecting the intersecting OCRs. This would cover more information and the accuracy would be assured by the downstream footprinting analysis and the integration of multiple filtering layers of information.

- The integration of multiple sources of data could also act as a limiting factor regarding the number of identified TFBS. Identifying the most restrictive data applied in this work and assessing alternative data sources could also increase the scope of the project.

Despite the mentioned limitations, we consider that we have generated a useful resource that, based on the integration of different omics in a comprehensive manner, has passed several filters and showed consistent results, which were also experimentally validated. Therefore, this resource can be helpful to make an *a priori* analysis to interrogate the WT mouse pancreas about transcriptional modules playing an important role in the regulation of acinar cell identity.

Although in this approach we have focused on WT mouse acinar pancreas, the aim of the project is to apply the explained methodology to other cell types, organisms and conditions. Based on the analysis exposed in figure 6, we believe that the results obtained in the present work can be partially extrapolated to human pancreas under homeostatic conditions. Therefore, building the transcriptional network underlying acinar cell identity in WT human could be useful to cross-validate the results obtained in mouse and establish parallelisms between the regulatory networks of both organisms. In fact, part of this work was driven in parallel for WT human making use of FAC-sorted pancreatic data, which gave us the possibility to directly perform the analyses at cell population level[88]. Global TF-TF networks were already generated for

acinar and ductal pancreatic cells (Supplementary figure 7), but additional analysis will be performed for the endocrine compartment in both human and mouse pancreas as quality control, focusing on beta cells-expressed TFs, as this is the predominant cell type in pancreatic islets.

In addition, future interrogation of acinar cell identity regulation in GEMMs displaying different tumorigenic conditions, such as KRAS-driven cancer models and KO mice for different acinar-expressed TFs studied in our group, can provide extremely valuable knowledge to the field of pancreatic cancer and, more specifically, of PDAC.

Methods

Datasets

High quality ATAC-seq datasets were downloaded from a publicly available resource on mouse epigenome (http://identifiers.org/ncbi/insdc.sra:SRP167062) using SRA toolkit fastq dump function (http://ncbi.github.io/sra-tools/fastq-dump.html). Two adult female (SRX4946168, SRX4946169) and two adult male (SRX4946145, SRX4946117) pancreas profiles were used for this project.

Processing of ATAC-seq data

Paired-end raw fastq files were analysed using the ENCODE ATAC-seq pipeline developed by Anshul Kundaje's laboratory[43]. Male and female replicates were analysed separately. The ENCODE pipeline allowed for an automated end-to-end quality control and processing of ATAC-seq data (Figure 12). Caper (Cromwell Assisted Pipeline ExecutoR) was used to run the pipeline from FASTQ to peak calling in an automated way (caper run [WDL script] -i [Input JSON file containing information of genomic data files, parameters and metadata]). Briefly, Cutadapt[89] v2.5 was used to find and remove the adapter sequences. Then, reads were mapped to reference genome (mm10, GRCm38, December 2011) using Bowtie2[90] v2.3.4.3 and the resulting SAM (Sequence Alignment Map) files were converted to BAM format using SAMtools[91] v1.9. Next, Sambamba[92] v0.6.6 was used to detect and remove reads unmapped, not primary alignment, duplicates and reads mapping to mitochondrial DNA (chrM). PCR read duplicates were removed by using Picard's MarkDuplicates[93]. Accessible regions were identified by peak calling using MACS2[94]. Consistent peaks between replicates were selected by applying an 0.05 IDR threshold.

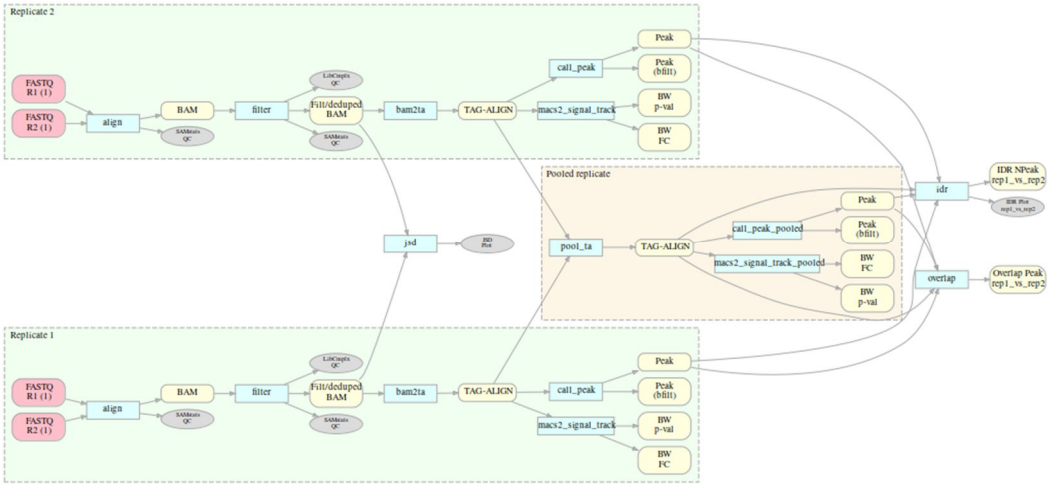


Figure 12. Steps followed in the analysis of ATAC-seq datasets. Identical end-to-end quality control from raw fastq files to peak calling was performed for both female and male replicates. Filtered BAM files and consistent OCRs obtained after this analysis constituted the basis for downstream footprinting analysis.

Quality control after processing showed robust results for the four analysed replicates. Regarding the library complexity, no bottlenecking was observed after checking the PCR bottlenecking coefficients (PBC1 and PBC2). PBC1 is the ratio of genomic locations with exactly one read pair over the genomic locations with at least one read pair. PBC2 is the ratio of genomic locations with exactly one read pair over the genomic locations with exactly two read pairs. All the replicates passed the threshold of TSS enrichment in OCRs for mm10 established as quality control.

Processing of ChIP-seq data

ChIPseq data from histone marks and transcription factors were analysed using Rubioseq pipeline[49] mounted in a Docker container. Reads were aligned to reference genome (mm10, GRCm38, December 2011) using Burrows-Wheeler Aligner (BWA). Duplicates were marked and removed using Picard. Peak calling was performed with MACS2 using *--nomodel --extsize 200 --gsize mm --broad-cutoff 0.01* argument for histone marks and *--nomodel --extsize 200 --gsize mm* for transcription factors.

Processing of RNA-seq data

RNAseq data was analysed using Nextpresso pipeline[95] mounted in a Docker container. Reads were aligned using Bowtie and TopHat2[96] aligners. Gene counts matrices were generated using HTseq-count[97] and normalization and differential expression analysis was carried out with DESeq2[98] package.

Processing of scRNA-seq data

Single cell RNA seq data was preprocessed using Cell Ranger software for Chromium 10X based data[99]. Sparse matrices were loaded into R for Seurat analysis[100] (cell QC, dimensionality reduction, normalization and graph-based clustering). Cell population annotation was based on the differential expression of canonical markers in each of the obtained clusters.

Selection and filtering of consensus OCR peaks

Peak files obtained for male and female replicates after ATAC-seq processing were merged to obtain a unique consensus peak file with robust OCRs. Both peak files were intersected using the HOMER software mergePeaks function. The "-d given" option was set to be conservative with the obtained common regions between both files. Instead of getting an average position between overlapping peaks, this option allowed to obtain a broader region comprising both overlapping peaks.

Peak files obtained from ChIP-seq analysed histone marks were used to filter and restrict the ATAC-seq consensus OCRs. Firstly, peak files from replicates for each histone mark were merged using HOMER's mergePeak function with -d given option.

Consensus peaks from H3K4me3 and H3K27me3 were intersected with the consensus ATAC-seq OCRs using bedtools intersect function. Histone mark locations, identified by ChIP-seq experiments, comprise a broader region than the real OCR identified by ATAC-seq, as the ChIP-seq peaks also include the histone position. The "-f " option was set as 0.5 to be restrictive enough to ensure a real overlap instead of the default 1bp to consider two peaks as common. With the -f 0.5 threshold it was required an overlap between the histone mark peak and a 50% of the OCR peak as a minimum to consider them as common peaks. The "-wa" option was set to keep the ATAC-seq peaks in the output file when an overlap occurs (bedtools intersect -a [ATAC-seq peaks] -b [Histone mark peaks] -wa -f 0.5). Consensus peaks from H3K27ac were divided into two subsets by intersecting them with TSS for GRCm38 downloaded from GENCODE, with +/- 1 kb around TSS. Bedtools intersect function was used with default parameters (bedtools intersect -a [H3K27ac peaks] -b [TSS +/- 1kb peaks] -wa). H3K27ac peaks overlapping in 1bp with TSS +/- 1kb regions were established as the promoter subset, while the non-overlapping peaks were established as the enhancer subset. Both subsets were intersected with the ATAC OCRs using the same procedure explained for H3K4me3 and H3K27me3.

Replicates of ChIP-seq identified coordinates for individual TFs were intersected with HOMER mergePeaks function to obtain the consensus peaks between replicates.

Peak annotation

Consensus peak files were annotated using HOMER annotatePeaks.pl function. mm10, GRCm38, December 2011 genome was used as reference genome and a gtf annotation file for mm10, GRCm38 extracted from the UCSC Genome Browser was used to annotate the peaks.

HOMER *de novo* motif analysis using the findMotifsGenome.pl function was performed to find enriched motifs in ChIP-seq consensus peaks for individual TFs. The PPM corresponding to the TF binding motif was used to annotate the peaks, again with HOMER annotatePeaks.pl, and differentiate regions matching and not matching the motif.

Footprinting analysis

TOBIAS ATACorrect function was used to correct the sequenced reads regarding Tn5 sequence bias. ScoreBigwig function was used to calculate footprint scores from cutsites across accessible regions. BINDetect module was used to estimate TF binding events from footprints and motif information. Default parameters were used in each step (0.001 p-value threshold for bound/unbound TF split). CreateNetwork module was used to model the interactions between bound TFs and their target genes to build regulatory networks.

Processing of motif information

TF binding motifs were mainly downloaded from CIS-BP[101] database for *Mus musculus*. Motifs for Ets2, Foxa3, Nfyb, Rbpjl, Sox9 and Tead2, known TFs with expression in pancreatic acinar cells, whose motifs were not present in CIS-BP, were downloaded from JASPAR CORE 2020[102].

The motif information included in the analysis was restricted to TFs expressed in pancreatic acinar cells. RNA-seq data was used to restrict the information to TFs with pancreatic expression levels greater than 1 RPKM. scRNA-seq data was used to restrict the motif information to pancreas-expressed TFs with expression in ≥ 1 acinar cell.

Visualization

Venn diagrams, barplots, piecharts, correlation plots and heatmaps were generated with RStudio v1.2.5019[103]. Network views were drawn with Cytoscape v3.8.2. Subnetworks with top nodes ranked by degree were drawn with *cytoHubba*[104] plugin in Cytoscape. Aggregated footprints were visualized with TOBIAS PlotAggregate function.

Statistical analysis

Pearson correlation coefficient was applied to measure the linear correlation between degree and betweenness centrality, as well as between these topological metrics and TF expression levels from RNA-seq analysis.

Statistical significance of the common results between networks and ChIP-seq analysis on individual TFs was assessed by random permutations of the ChIP-seq regions. These permuted regions were annotated, and the identified target genes were compared with those identified in the corresponding network. Next, the number of common genes was compared with the enrichment obtained from the comparison of the network and the real ChIP-seq data. Pearson's chi-squared test was used to determine the statistical significance of the observed consistency of the results relative to the expected enrichment[105].

All statistical analyses were performed using R[106].

Code availability

For detailed information of the code used to analyse and represent the results obtained see https://github.com/PabloPerez5/TFM.

References

1. Macdonald, R. J., Swift, G. H. & Real, F. X. Chapter 1 - Transcriptional Control of Acinar Development and Homeostasis. **97**, 1–40 (2010).

2. Hezel, A. F., Kimmelman, A. C., Stanger, B. Z., Bardeesy, N. & DePinho, R. A. Genetics and biology of pancreatic ductal adenocarcinoma. *Genes and Development* **20**, 1218–1249 (2006).

3. Pour, P. M., Pandey, K. K. & Batra, S. K. What is the origin of pancreatic adenocarcinoma? *Molecular Cancer* **2**, 1–10 (2003).

4. Kamisawa, T., Wood, L. D., Itoi, T. & Takaori, K. Pancreatic cancer. *The Lancet* **388**, 73–85 (2016).

5. Rahib, L. *et al.* Projecting cancer incidence and deaths to 2030: The unexpected burden of thyroid, liver, and pancreas cancers in the united states. *Cancer Research* **74**, 2913–2921 (2014).

6. Kyriazis, A. A., Kyriazis, A. P., Sternberg, C. N., Sloane, N. H. & Loveless, J. D. Morphological, Biological, Biochemical, and Karyotypic Characteristics of Human Pancreatic Ductal Adenocarcinoma Capan-2 in Tissue Culture and the Nude Mouse. *Cancer Res.* **46**, 5810–5815 (1986).

7. Aguirre, A. J. *et al.* Activated Kras and Ink4a/Arf deficiency cooperate to produce metastatic pancreatic ductal adenocarcinoma. *Genes Dev.* **17**, 3112–3126 (2003).

8. Hingorani, S. R. *et al.* Preinvasive and invasive ductal pancreatic cancer and its early detection in the mouse. *Cancer Cell* **4**, 437–450 (2003).

9. Habbe, N. *et al.* Spontaneous induction of murine pancreatic intraepithelial neoplasia (mPanIN) by acinar cell targeting of oncogenic Kras in adult mice. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 18913–18918 (2008).

10. Friedlander, S. Y. G. *et al.* Context-Dependent Transformation of Adult Pancreatic Cells by Oncogenic K-Ras. *Cancer Cell* **16**, 379–389 (2009).

11. Bailey, J. M. *et al.* P53 mutations cooperate with oncogenic Kras to promote adenocarcinoma from pancreatic ductal cells. *Oncogene* **35**, 4282–4288 (2016).

12. Brembeck, F. H. *et al.* The mutant K-ras oncogene causes pancreatic periductal lymphocytic infiltration and gastric mucous neck cell hyperplasia in transgenic mice. *Cancer Res.* **63**, 2005–2009 (2003).

13. Kopp, J. L. *et al.* Identification of Sox9-Dependent Acinar-to-Ductal Reprogramming as the Principal Mechanism for Initiation of Pancreatic Ductal Adenocarcinoma. *Cancer Cell* **22**, 737–750 (2012).

14. Slack, J. M. W. Metaplasia and transdifferentiation: From pure biology to the clinic. *Nature Reviews Molecular Cell Biology* **8**, 369–378 (2007).

15. Stanger, B. Z. & Hebrok, M. Control of cell identity in pancreas development and

regeneration. *Gastroenterology* **144**, 1170–1179 (2013).

16. Wong, C. H., Li, Y. J. & Chen, Y. C. Therapeutic potential of targeting acinar cell reprogramming in pancreatic cancer. *World Journal of Gastroenterology* **22**, 7046–7057 (2016).

17. Martinelli, P. *et al.* The acinar regulator Gata6 suppresses KRasG12V-driven pancreatic tumorigenesis in mice. *Gut* **65**, 476–486 (2016).

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

19. Segal, E., Yelensky, R. & Koller, D. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics* **19**, i273-282 (2003).

20. Kornberg, R. D. Chromatin Structure: A Repeating Unit of Histones and DNA. *Science* **184**, 868–871 (1974).

21. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).

22. Roeder, R. The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem. Sci.* **21**, 327–335 (1996).

23. Nikolov, D. B. & Burley, S. K. RNA polymerase II transcription initiation: A structural view. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 15–22 (1997).

24. Volpe, T. A. *et al.* Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* **297**, 1833–1837 (2002).

25. Ting, D. T. *et al.* Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. *Science* **331**, 593–596 (2011).

26. Volpe, T. & Martienssen, R. A. RNA interference and heterochromatin assembly. *Cold Spring Harb. Perspect. Biol.* **3**, 1–11 (2011).

27. Lee, C. K., Shibata, Y., Rao, B., Strahl, B. D. & Lieb, J. D. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nat. Genet.* **36**, 900–905 (2004).

28. Bernstein, B. E., Meissner, A. & Lander, E. S. The Mammalian Epigenome. *Cell* **128**, 669–681 (2007).

29. Quina, A. S., Buschbeck, M. & Di Croce, L. Chromatin structure and epigenetics. *Biochem. Pharmacol.* **72**, 1563–1569 (2006).

30. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).

31. Mullen, A. C. *et al.* Master transcription factors determine cell-type-specific responses to TGF-β signaling. *Cell* **147**, 565–576 (2011).

32. Yang, C. C. *et al.* Inferring condition-specific targets of human TF-TF complexes

using ChIP-seq data. *BMC Genomics* **18**, 1–10 (2017).

33. Kitano, H. Systems biology: A brief overview. *Science* **295**, 1662–1664 (2002).

34. Brazhnik, P., De La Fuente, A. & Mendes, P. Gene networks: How to put the function in genomics. *Trends in Biotechnology* **20**, 467–472 (2002).

35. Davidson, E. H. & Erwin, D. H. Gene regulatory networks and the evolution of animal body plans. *Science* **311**, 796–797 (2006).

36. Mercatelli, D., Scalambra, L., Triboli, L., Ray, F. & Giorgi, F. M. Gene regulatory network inference resources: A practical overview. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* **1863,** 194430 (2020).

37. Margolin, A. A. *et al.* Reverse engineering cellular networks. *Nat. Protoc.* **1**, 662–671 (2006).

38. Zhu, J. *et al.* Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat. Genet.* **40**, 854–861 (2008).

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

40. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **2015**, 21.29.1-21.29.9 (2015).

41. Galas, D. J. & Schmitz, A. DNAase footprinting a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.* **5**, 3157–3170 (1978).

42. Liu, C. *et al.* An ATAC-seq atlas of chromatin accessibility in mouse tissues. *Sci. Data* **6**, 65 (2019).

43. Anshul, K., Nathan, B., Daniel, K., Chuan Sheng, F. & Jin wook Lee. ENCODE ATAC-seq pipeline. *ENCODE-DCC, GitHub repository* https://github.com/ENCODE-DCC/atac-seq-pipeline (2016).

44. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5**, 1752–1779 (2011).

45. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38**, 576–589 (2010).

46. Liang, G. *et al.* Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. *Proc. Natl. Acad. Sci.* **101**, 7357–7362 (2004).

47. Pradeepa, M. M. Causal role of histone acetylations in enhancer function. *Transcription* **8**, 40–47 (2017).

48. Barski, A. *et al.* High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129**, 823–837 (2007).

49. Rubio-Camarillo, M. *et al.* RUbioSeq+: A multiplatform application that executes parallelized pipelines to analyse next-generation sequencing data. *Comput. Methods Programs Biomed.* **138**, 73–81 (2017).

50. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

51. Bentsen, M. *et al.* ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nat. Commun.* **11**, 4267 (2020).

52. Koohy, H., Down, T. A. & Hubbard, T. J. Chromatin Accessibility Data Sets Show Bias Due to Sequence Specificity of the DNase I Enzyme. *PLoS One* **8**, e69853 (2013).

53. Karabacak Calviello, A., Hirsekorn, A., Wurmus, R., Yusuf, D. & Ohler, U. Reproducible inference of transcription factor footprints in ATAC-seq and DNase-seq datasets using protocol-specific bias modeling. *Genome Biol.* **20**, 1–13 (2019).

54. He, H. H. *et al.* Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat. Methods* **11**, 73–78 (2014).

55. Li, Z. *et al.* Identification of transcription factor binding sites using ATAC-seq. *Genome Biol.* **20**, 45 (2019).

56. Siddharthan, R. Dinucleotide weight matrices for predicting transcription factor binding sites: Generalizing the position weight matrix. *PLoS One* **5**, e9722 (2010).

57. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* **45**, 580–585 (2013).

58. Cobo, I. *et al.* Transcriptional regulation by NR5A2 links differentiation and inflammation in the pancreas. *Nature* **554**, 533–537 (2018).

59. Shannon, P. *et al.* Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

60. Direnzo, D. *et al.* Induced Mist1 expression promotes remodeling of mouse pancreatic acinar cells. *Gastroenterology* **143**, 469–480 (2012).

61. Hale, M. A. *et al.* The nuclear hormone receptor family member NR5A2 controls aspects of multipotent progenitor cell formation and acinar differentiation during pancreatic organogenesis. *Dev.* **141**, 3123–3133 (2014).

62. Martinelli, P. *et al.* GATA6 regulates EMT and tumour dissemination, and is a marker of response to adjuvant chemotherapy in pancreatic cancer. *Gut* **66**, 1665–1676 (2017).

63. Gong, Y. *et al.* GATA4 inhibits cell differentiation and proliferation in pancreatic cancer. *PLoS One* **13,** e0202449 (2018).

64. Carrasco, M., Delgado, I., Soria, B., Martín, F. & Rojas, A. GATA4 and GATA6 control mouse pancreas organogenesis. *J. Clin. Invest.* **122**, 3504–3515 (2012).

65. Xuan, S. *et al.* Pancreas-specific deletion of mouse Gata4 and Gata6 causes pancreatic agenesis. *J. Clin. Invest.* **122**, 3516–3528 (2012).

66. Xuan, S. & Sussel, L. GATA4 and GATA6 regulate pancreatic endoderm identity through inhibition of hedgehog signaling. *Dev.* **143**, 780–786 (2016).

67. Villamayor, L., Cano, D. A. & Rojas, A. GATA factors in pancreas development and disease. *IUBMB Life* **72**, 80–88 (2020).

68. Kumazoe, M. *et al.* FOXO3 is essential for CD44 expression in pancreatic cancer cells. *Oncogene* **36**, 2643–2654 (2017).

69. Costello, L. C., Zou, J., Desouki, M. M. & Franklin, R. B. Evidence for changes in RREB-1, ZIP3, and zinc in the early development of pancreatic adenocarcinoma. *J. Gastrointest. Cancer* **43**, 570–578 (2012).

70. Shin, S. *et al.* Activator protein-1 has an essential role in pancreatic cancer cells and is regulated by a novel Akt-mediated mechanism. *Mol. Cancer Res.* **7**, 745–754 (2009).

71. Prévot, P. P. *et al.* Role of the ductal transcription factors HNF6 and Sox9 in pancreatic acinar-to-ductal metaplasia. *Gut* **61**, 1723–1732 (2012).

72. Park, J. *et al.* YAP and AP-1 cooperate to initiate pancreatic cancer development from ductal cells in Mice. *Cancer Res.* **80**, 4768–4779 (2020).

73. Freeman, L. C., Roeder, D. & Mulholland, R. R. Centrality in social networks: ii. experimental results. *Soc. Networks* **2**, 119–141 (1979).

74. Newman, M. E. J. The structure and function of complex networks. *SIAM Review* **45**, 167–256 (2003).

75. Colizza, V., Flammini, A., Serrano, M. A. & Vespignani, A. Detecting rich-club ordering in complex networks. *Nat. Phys.* **2**, 110–115 (2006).

76. Kintali, S. Betweenness Centrality : Algorithms and Lower Bounds. (2008).

77. Mirzasoleiman, B. & Jalili, M. Failure tolerance of motif structure in biological networks. *PLoS One* **6**, e20512 (2011).

78. Safe, S. & Abdelrahim, M. Sp transcription factor family and its role in cancer. *Eur. J. Cancer* **41**, 2438–2448 (2005).

79. Fernandez-Zapico, M. E. *et al.* A functional family-wide screening of SP/KLF proteins identifies a subset of suppressors of KRAS-mediated cell growth. *Biochem. J.* **435**, 529–537 (2011).

80. Jiang, W., Cui, J., Xie, D. & Wang, L. Sp/KLF Family and Tumor Angiogenesis in Pancreatic Cancer. *Curr. Pharm. Des.* **18**, 2420–2431 (2012).

81. Tetreault, M. P., Yang, Y. & Katz, J. P. Krüppel-like factors in cancer. *Nature Reviews Cancer* **13**, 701–713 (2013).

82. Zhang, D. *et al.* KLF2 is downregulated in pancreatic ductal adenocarcinoma and inhibits the growth and migration of cancer cells. *Tumor Biol.* **37**, 3425–3431

(2016).

83. Zhu, Z. *et al.* Krüppel-Like Factor 4 Inhibits Pancreatic Cancer Epithelial-to-Mesenchymal Transition and Metastasis by Down-Regulating Caveolin-1 Expression. *Cell. Physiol. Biochem.* **46**, 238–252 (2018).

84. He, P., Yang, J. W., Yang, V. W. & Bialkowska, A. B. Krüppel-like Factor 5, Increased in Pancreatic Ductal Adenocarcinoma, Promotes Proliferation, Acinar-to-Ductal Metaplasia, Pancreatic Intraepithelial Neoplasia, and Tumor Growth in Mice. *Gastroenterology* **154**, 1494-1508.e13 (2018).

85. Le, D. D. *et al.* Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E3702–E3711 (2018).

86. Barne, S. L., Belliveau, N. M., Ireland, W. T., Kinney, J. B. & Phillips, R. Mapping DNA sequence to transcription factor binding energy in vivo. *PLoS Comput. Biol.* **15**, e1006226 (2019).

87. Liu, J., Shively, C. A. & Mitra, R. D. Quantitative analysis of transcription factor binding and expression using calling cards reporter arrays. *Nucleic Acids Res.* **48**, e50–e50 (2020).

88. Herzenberg, L. A., De Rosa, S. C. & Herzenberg, L. A. Monoclonal antibodies and the FACS: Complementary tools for immunobiology and medicine. *Immunology Today* **21** 383–390 (2000).

89. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).

90. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

91. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

92. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: Fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).

93. 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. *Broad Institute, GitHub repository* https://github.com/broadinstitute/picard (2019).

94. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).

95. Graña, O., Rubio-Camarillo, M., Fdez-Riverola, F., Pisano, D. G. & Glez-Peña, D. Nextpresso: Next Generation Sequencing Expression Analysis Pipeline. *Curr. Bioinform.* **13**, 583–591 (2017).

96. Kim, D. *et al.* TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, 1–13 (2013).

97. Anders, S., Pyl, P. T. & Huber, W. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).

98. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

99. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 1–12 (2017).

100. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).

101. Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).

102. Fornes, O. *et al.* JASPAR 2020: Update of the open-Access database of transcription factor binding profiles. *Nucleic Acids Res.* **48**, D87–D92 (2020).

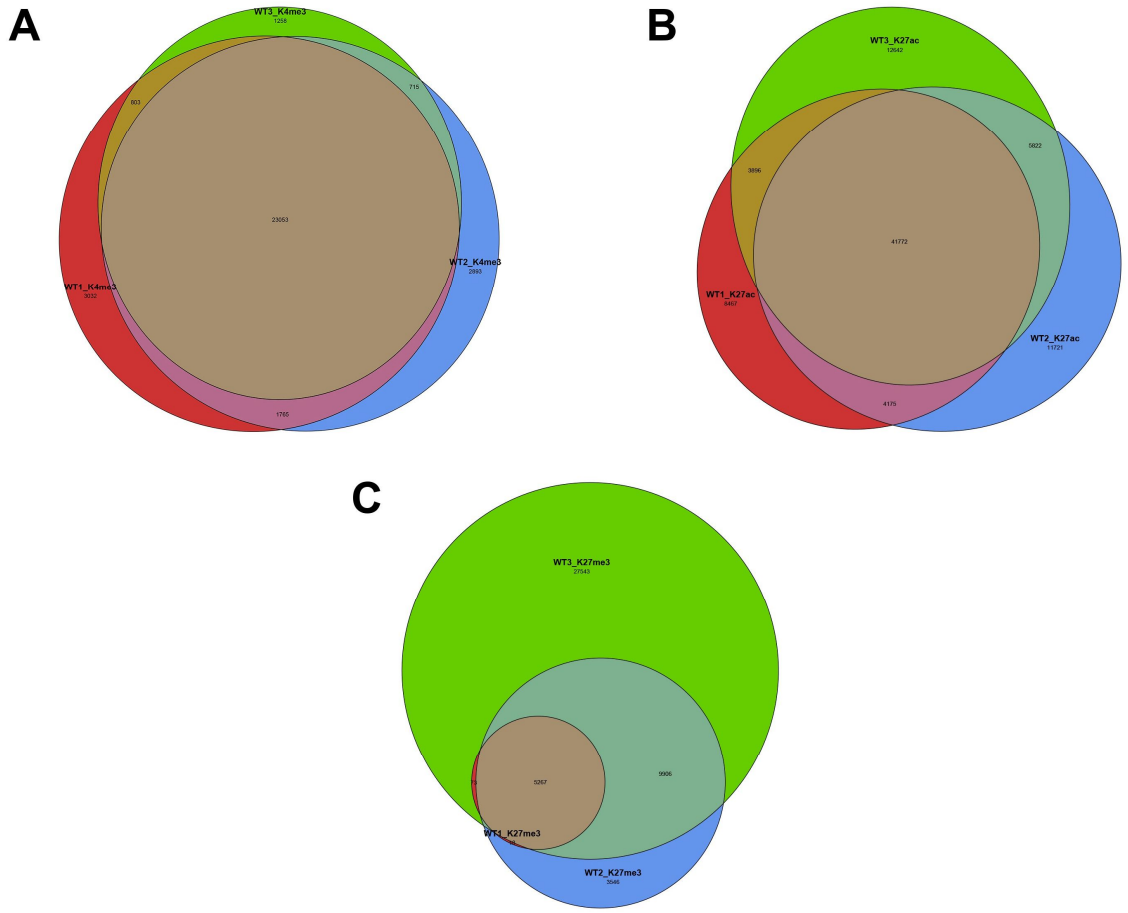
103. RStudio Team (2020). RStudio: Integrated Development for R. *RStudio, PBC, Boston, MA.* https://www.rstudio.com/.

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

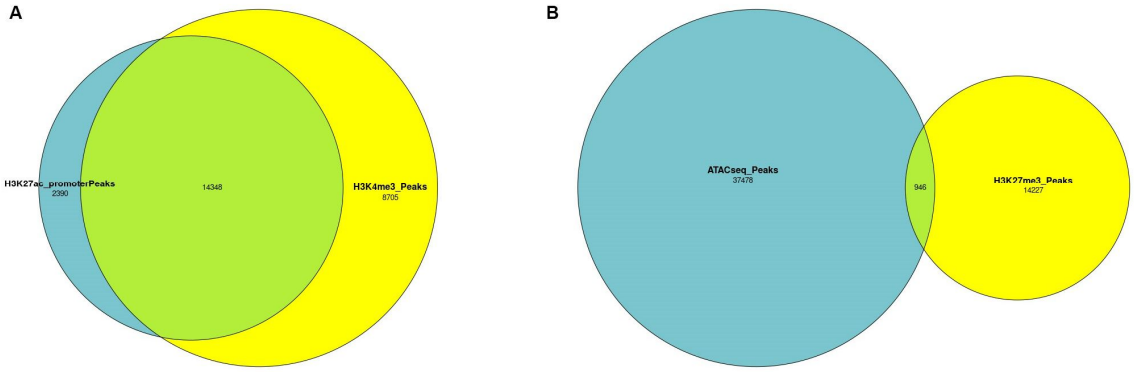
105. Pearson, K. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling . *London, Edinburgh, Dublin Philos. Mag. J. Sci.* **50**, 157–175 (1900).

106. R Core Team (2020). R: The R Project for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria.* https://www.r-project.org/.

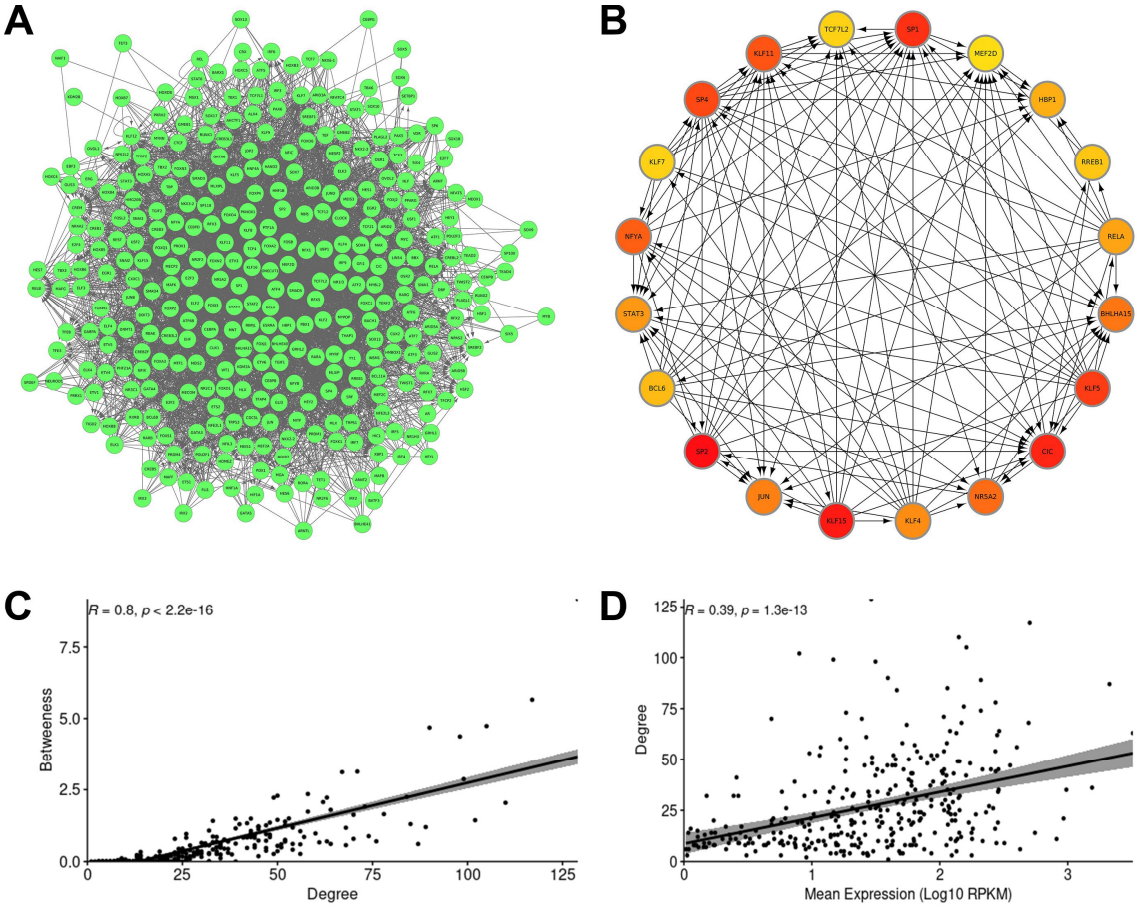
Supplementary Information



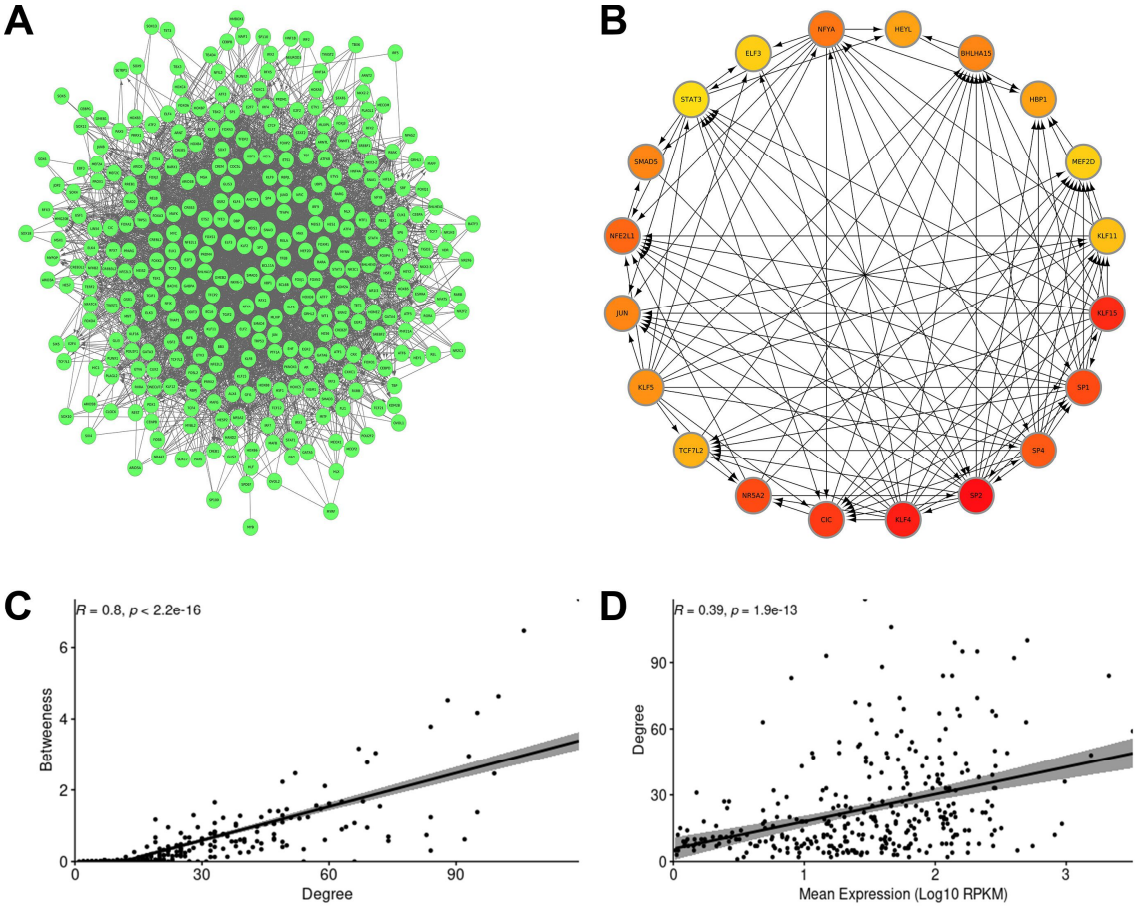
Supplementary Figure 1. Histone marks consensus peaks. Venn diagrams showing the peak regions identified by ChIP-seq experiments for different epigenetic modifications. **A)** For H3K4me3, there are 23053 common peaks identified by the three replicates. **B)** 41772 consensus regions between the three replicates were identified for H3K27ac. **C)** In the case of H3K27me3, considering that the first replicate has few peaks compared with the others and almost all of them are overlapping, it was excluded and the 15173 common peaks between the other two replicates were took as the consensus peak set.



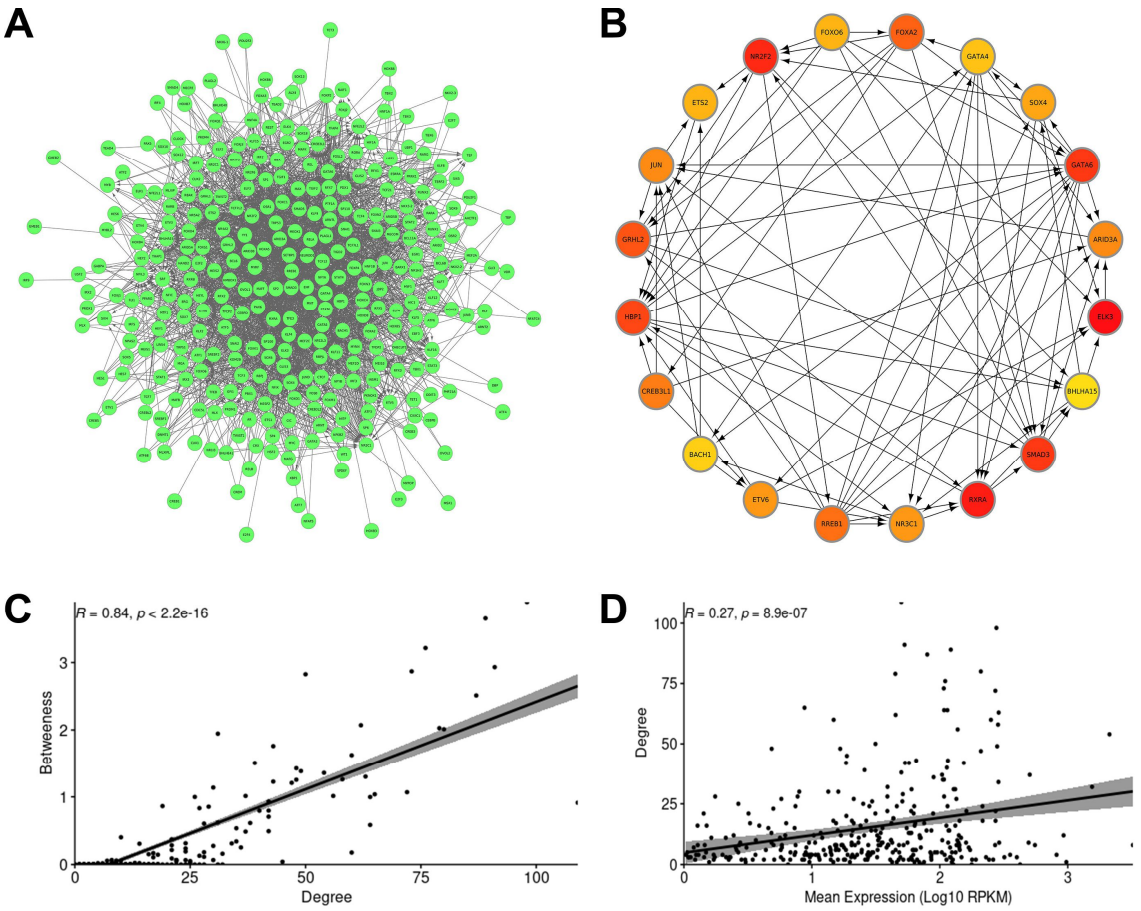
Supplementary Figure 2. Peak sets consistency verification. A) Promoter regions identified by H3K4me3 and H3K27ac (promoter subset). Most of H3K27ac peaks overlaps with H3K4me3 peaks. **B)** Overlap between the OCRs identified by ATAC-seq and repressed regions marked by the H3K27me3 epigenetic modification. A minimum intersection is obtained.



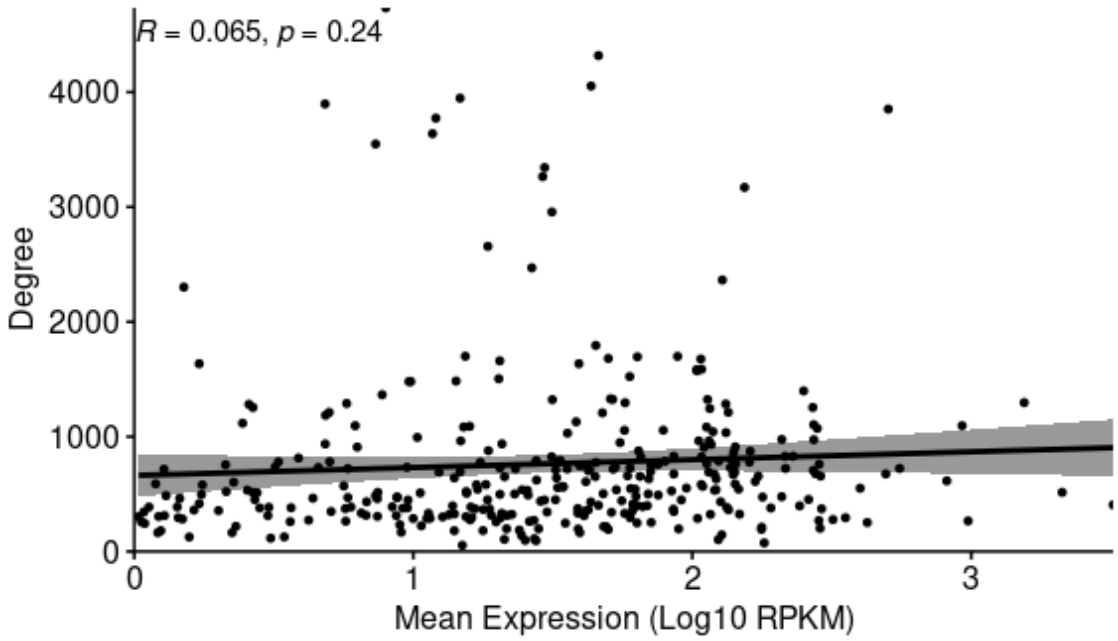
Supplementary Figure 3. Construction and topological analysis of TF-TF network (H3K4me3-filtered peak set). A) Global view of the TF-TF network for the set of peaks restricted to promoters by H3K4me3. **B)** Top 20 nodes (TFs) of the network ranked by degree. **C)** Correlation between degree and betweenness centrality measures for the H3K4me3-filtered network. A strong correlation was found between both local and global methods. **D)** Correlation between nodes degree and expression levels of TFs. A weak correlation was found.



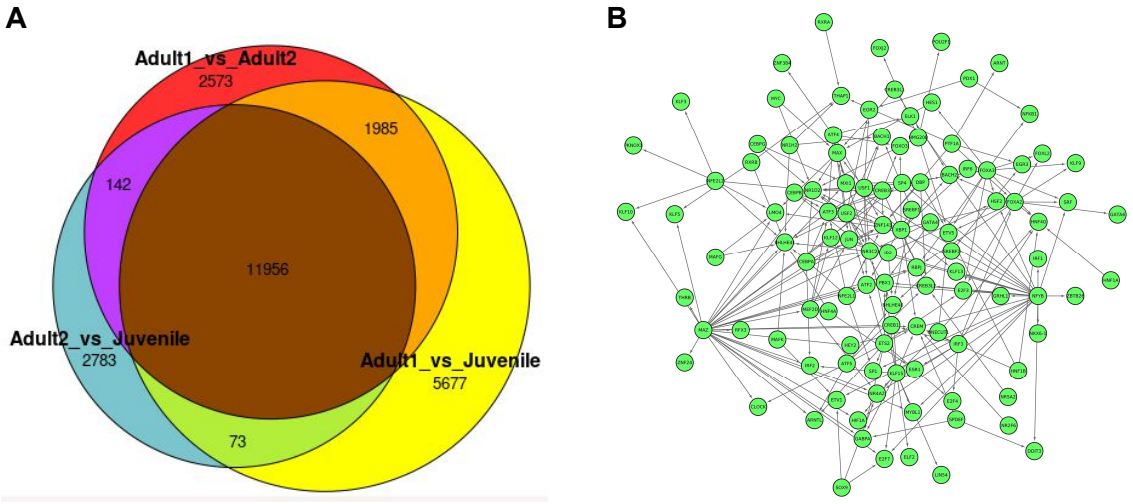
Supplementary Figure 4. Construction and topological analysis of TF-TF network (H3K27ac promoter subset-filtered peak set). A) Global view of the TF-TF network for the set of peaks restricted to promoters by H3K27ac. **B)** Top 20 nodes (TFs) of the network ranked by degree. **C)** Correlation between degree and betweenness centrality measures for the H3K27ac promoter subset-filtered network. A strong correlation was found between both local and global methods. **D)** Correlation nodes degree and expression levels of TFs. A weak correlation was found.



Supplementary Figure 5. Construction and topological analysis of TF-TF network (H3K27ac enhancer subset-filtered peak set). A) Global view of the TF-TF network for the set of peaks restricted to enhancers by H3K27ac. **B)** Top 20 nodes (TFs) of the network ranked by degree. **C)** Correlation between degree and betweenness centrality measures for the H3K27ac enhancer subset-filtered network. A strong correlation was found between both local and global methods. **D)** Correlation between nodes degree and expression levels of TFs. A weak correlation was found.



Supplementary Figure 6. Correlation between the degree of TF nodes and mean expression levels in the global network built for all acinar-expressed TFs and their target genes. No correlation was found.



Supplementary Figure 7. Parallel analysis performed for human acinar pancreas. A) OCRs identified by ATAC-seq in WT human acinar pancreas. 11956 consensus regions were identified between replicates. **B)** TF-TF network built from the 11956 consensus OCRs identified in human acinar pancreas under homeostatic conditions, consisting of 110 nodes (TF) and 242 edges (interactions).

Transcription factor	Average expression
Xbp1	3224,526308265
Bhlha15	2117,0835867338
Nme2	1676,2067027481
Rbpjl	1547,5161185468
Atf4	974,2099002556
Atf5	926,9699547841
Tead2	817,1204197394
Atf6	553,6087661701
Son	545,9655962968
Klf15	504,1595133485
Jund	493,9339829912
Preb	467,2507425435
Kcmf1	460,7877138226
Pa2g4	427,8411133771
Glyr1	427,6188023606
Cenpb	423,8887302389
Nfe2l1	399,7284173814
Gtf2i	357,191192227
Cxxc1	354,1297566215
Cux1	319,2360390111
Tsc22d1	305,8274320829
Tef	290,717258602
Nr3c1	288,3094275963
Klf9	287,1549493714
Nr1d2	285,9913122555
Gata4	285,0187012837
Irf6	284,2570659822
Ptf1a	280,5579125471
Rxra	276,7830786548
Mbnl2	276,3122628736
Nfic	272,7066766024
Stat3	272,4564463418
Creb3l1	271,5448377125
Tfdp2	270,3198562047
Srebf1	261,1951981888
Ets2	250,7319207906
Dbp	242,5844532446
Cxxc5	235,9642600382
Usf2	229,6053672063
Chchd3	229,507090639
Drap1	227,7459621267
Nr2f6	215,9091003994
Mlxip	215,2683237544
Nr5a2	209,259546813
Hbp1	209,1492884816
Purb	208,1692629418
Ski	198,9572394732
Nr1d1	198,7511440913
Aebp1	196,7322470996
Ubp1	191,1920450102

Transcription factor	Average expression
Deaf1	188,7294904037
Ppard	181,448328333
Kdm2a	181,111201324
Nfat5	178,6499605608
Cebpg	177,3238040221
Cebpa	176,9674109971
Epas1	175,3177519375
Mbd2	173,0352093183
Cggbp1	172,9839539462
Creb3	172,3834590697
Stat6	168,1824840878
Cebpz	167,8869317995
Baz2a	167,4051861618
Gpbp1	165,4124227595
Sp1	161,2911401825
Meis2	160,2990907802
Nfib	160,2333664851
Nfx1	156,6802720181
Rela	153,9129851835
Ash1l	148,8350627154
Mef2d	147,6892814619
Mxd4	147,4266308338
Adnp	146,9725942487
Hnf4a	146,6085129266
Prdm2	143,226649482
Mlx	143,1340664599
Foxp4	142,5768717475
Ahctf1	142,4538485431
Rxrb	141,9307584946
Gpbp1l1	141,0357122983
Cic	140,7384571155
Myc	140,1365480884
Klf6	139,8394678731
Bptf	139,7831472151
Bach1	137,8030206875
Irf3	136,9386646741
Mxi1	135,1111076508
Hif1a	134,9045478783
Smad5	134,6394426612
Sp3	133,8779044269
Clock	132,8726913994
Cdc5l	132,07321345
Tcf4	131,8396451176
Tead1	129,7358286923
Gatad2a	128,4493152467
Ctcf	128,2050977072
Crebzf	127,5420718557
Srebf2	126,4076907638
Klf13	125,5254618471
Usf1	124,2997115701

Supplementary Table 1. Top 100 most expressed (RPKM) transcription factors in WT mouse pancreas.

Xbp1	Bach1	Esrra	Trp53	Hsf2	Etv1	Twist1
Bhlha15	Irf3	Stat1	Pou2f1	Nkx2-2	E2f3	Foxc1
Rbpjl	Hif1a	Hes6	Kdm2b	Tcf7	Glis3	Nr1i3
Atf4	Smad5	E2f4	Tgif1	Irf7	Tgif2	Egr2
Atf5	Clock	Crebl2	Nfya	Elf4	Irf4	Rel
Tead2	Cdc5l	Foxk1	Glis2	Elk4	Naif1	Foxs1
Atf6	Tcf4	Rfx7	Nfil3	Gmeb1	Ebf3	Alx4
Klf15	Ctcf	Foxp2	Mnt	Tcf7l1	Arid3a	Creb5
Jund	Crebzf	Nfyb	Dnmt1	E2f2	Atf3	Hoxb6
Cenpb	Srebf2	Gmeb2	Rfx1	Tfap4	Gli3	Tbx6
Nfe2l1	Usf1	Hnf1b	Rarg	Bcl6	Nfatc4	Foxq1
Cxxc1	Tet3	Hsf1	Junb	Relb	Klf5	Hoxd8
Cux1	Smad3	Irf2	Mef2a	Crem	Snai3	Fosb
Tef	Smad4	Phf21a	Mynn	Pparg	Hey1	Tead4
Nr3c1	Arnt	Gfi1	Meis3	Hlf	Pax6	E2f7
Klf9	Atf6b	Hnf1a	Tfeb	Mitf	Tbx2	Etv4
Gata4	Foxa3	Mecom	Nr4a2	Osr1	Klf8	Hoxc4
Irf6	Foxo4	Bbx	Hes1	Hic1	Vdr	Msx1
Ptf1a	Atf1	Mtf1	Pknox1	Arid3b	Tet1	Sp6
Rxra	Tcf12	Etv5	Plagl2	Sox7	Nkx2-3	Irx3
Nfic	Nfix	Nr1h3	Tfcp2	Bcl11a	Sox17	Hey2
Stat3	Jun	Elf3	Klf11	Irf5	Prdm1	Pax5
Creb3l1	Yy1	Foxn2	Rara	Rest	Runx1	Runx2
Tfdp2	Spdef	Stat2	Hmbox1	Maff	Foxm1	Arnt2
Srebf1	Rbpj	Terf2	Klf16	Hes7	Npas2	Hoxc5
Ets2	Foxa2	Foxj2	Ddit3	Rora	Snai1	Gata3
Dbp	Mga	Heyl	Sp2	Elk1	Mesp2	Crx
Usf2	Nfe2l2	Nr2f2	Lin54	Wt1	Sox5	Irx2
Nr2f6	Etv6	Meis1	Cebpd	Homez	Hoxa5	Prrx2
Mlxip	Max	Ets1	Mlxipl	Foxo6	Ar	Tbx1
Nr5a2	Rreb1	Sox6	Rfx5	Atf7	Nkx6-1	Hoxb8
Hbp1	Foxo1	Tcf3	Sp100	Sp4	Bcl6b	Batf3
Ubp1	Pbx1	Elk3	Ovol2	Fli1	Hand2	Nkx3-2
Kdm2a	Ehf	Tigd2	Sox18	Mypop	Stat4	Twist2
Nfat5	Gabpa	Six5	Rbak	Jdp2	Insm1	Sox10
Cebpg	Mafk	Sox13	Prrx1	Six4	Klf7	
Cebpa	Mecp2	Sox12	Foxn3	Trps1	Tcf21	
Creb3	Srf	Etv3	Mef2c	Nfe2l3	Tbx3	
Stat6	Elf2	Irf9	Setbp1	Grhl1	Pou2f2	
Sp1	Sox9	Klf4	Tcf7l2	Bhlhe41	Rarb	
Meis2	Creb3l2	Sox4	Arid5b	Sp110	Hoxb4	
Rela	Atf2	Egr1	Ovol1	Klf12	Neurod1	
Mef2d	Arid2	Grhl2	Mafb	Onecut1	Mybl2	
Hnf4a	Bhlhe40	Klf2	Mafg	Prox1	Gata5	
Mlx	Gata6	Fosl2	Pdx1	Barx1	Snai2	
Foxp4	Foxj3	Arntl	Arid5a	Hlx	Myb	
Ahctf1	Plagl1	Prdm4	Tbp	Rfx2	Hoxb3	
Rxrb	Creb1	Cebpb	Rfx3	Foxj1	Hoxb7	
Cic	Hmg20b	Nr2c1	Cux2	Meox1	Hoxb5	
Myc	Tfe3	Nfkb2	Thap1	Erg	Osr2	

Supplementary Table 2. List of TFs selected as expressed in pancreatic acinar cells.